University of Zurich[UZH]

Burkhard Stiller, Andri Lareida, Lisa Kristiana, Thomas Bocek, Patrick Poullie, Bruno Bastos Rodrigues, and Corinna Schmitt (Eds.)

Internet Economics XI

TECHNICAL REPORT – No. IFI-2017.01

January 2017

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

B. Stiller et al. (Eds.): Internet Economics XI
Technical Report No. IFI-2017.01, January 2017
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.csg.uzh.ch/

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 2016 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 XI" discusses a number of selected topics in the area of Internet Economics. the first talk "Economic Benefits of Automated Vehicle Systems" discusses vehicle communication, which is an upcoming topic in the area of vehicle connectivity, with special focus on economic benefits. Talk two "Economics of Radio Towers" discusses the various economic aspects of todays deployed radio towers establishing connectivity everywhere. Talk three "Comparing Blockchains" discusses the hype of the crypto currency market. Talk four on "Emerging Pricing Models for Cloud Services" presents a solution for comparing the impact of different pricing models and compiled prices from five different cloud service providers to determine underlying pricing patterns and to understand the price differences across different cloud service providers. Talk five entitled "Different Impacts of the New Swiss Law on Monitoring of Postal and Telecommunication Traffic" highlights the main areas costs are increasing in the future, whereby small and medium sized companies suffer the most due to high investment costs. Talk six entitled "The Breaching of Cloud SLAs: Why it Happens and How to Prove it" focus on the current Cloud development and its consequences and challenges. Talk seven entitled "Microtransactions with the Lightning Network" takes up the hype of crypto currencies and introduces the Lightning Network addressing the scalability problem of Bitcoin by using off-blockchain techniques. Talk eight entitled

"Economical Impact of SDN and NFV for Telecom Operators" presents examples for SDN and NFV discussing their benefits and drawbacks related to the transition towards a software-defined and virtualized networking infrastructure. Talk nine entitled "Challenges in 5G: Social, Technology, and Economic Perspectives" investigates key challenges of 5G mobile communications and reveals key technologies that have been suggested by academia as well as representatives of the industry, and implications on economical and social perspectives. Talk ten entitled "Comparison of IoT Business Models" discusses common business models focusing on the market field of Internet-of-Things (IoT) applying them to a possible company to compare them in order to recommend one.

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 Andri Lareida, Lisa Kristiana, Thomas Bocek, Patrick Poullie, Bruno Bastos Rodrigues, Corinna Schmitt, and Burkhard Stiller. In particular, many thanks are addressed to Corinna Schmitt for her 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 2017

Contents

Chapter 1

Economic benefits of Automated Vehicle Systems

Andreas Milz, Yves Steiner

This paper studies the economic impact of the proliferation of automated vehicles for individual and public transportation. This work covers the topic of automated vehicle technology from an economic perspective. Technical details are only covered where it is necessary to understand the economic impact of automated vehicle technology. This study leads to a main question: „What economic benefits does the proliferation of automated vehicle systems for individual and private transportation yield and what challenges and risks are to be expected?"Starting from this point we first give an overview on technologies in the field of automated vehicles currently available or envisioned in the future. The potential benefits and risks of these technologies are summarized and discussed scientifically. The basis for our analysis are four scenarios for the adoption of automated vehicle technology, referring to the future ownership and usage of such vehicles. We cover two approaches how benefits and risks of automated vehicles can be measured empirically, namely classical cost/benefit analysis and approaches that try to incorporate non-monetizable and human factors. In the end, we compare the four scenarios of adoption in regard of their individual economic benefits and risks.

Contents

1.1 Introduction

Automated vehicle systems summarize a huge number of different vehicle types. To make a valid argument, the scope of this paper has to be narrowed down first. The broad topic of automated vehicle systems consists of ground vehicles, including every automated moving vehicle on the ground (e.g. autonomous tractors [1], cars, freight transport and automated vehicles in other fields like warehouses, freight ports). Ground vehicles, furthermore, can be divided into three subgroups: automated guided vehicles, trucks and cars. The last subcategory, cars, is of interest in this paper and subject of our economical analysis. The two subcategories which are not further addressed in our work, the automated guided vehicles, are robots that navigate by markers, wires in the floor, or uses vision, magnets, or lasers for orientation purpose and trucks with convoy systems or other automation technologies.
Cars, in this case autonomous cars, have different levels of automation which are the result of its system capabilities. The society of automotive engineers (SAE), which is an organisation providing standards and best practices for industries, released the norm J3016. It describes the classification and definition of automation found in autonomous vehicles, placing them in one of six levels of automation, depending on their capabilities. The six levels of automation range from level zero, with no automation going on, up to level five, where we find so-called full automation. This classification is based on four criterias. The considered criterias, execution of steering and acceleration or deceleration, monitoring of driving environment, fallback performance of dynamic driving task and system capability, are used to describe and classify the vehicles abilities. The six levels and four criterias give rise to the table in figure 1.1. The table indicates that, the higher the level of automation, the more functions are in the responsibility of the automated system. Vehicles at levels zero to two are vehicles, where its human driver is responsible for monitoring its environment. Functions such as cruise control, traffic jam assists or lane keeping assistance fall into these levels. For vehicles at levels three to five, the monitoring of the environment is managed by the vehicle system.
Consequently, in levels three to five we find vehicles equipped with systems capable of monitoring their surroundings. Examples for vehicle systems residing in level three encompass automated parking systems or a traffic jam pilots. Those systems show limited capabilities to take over the vehicle but necessitate a human driver to be fully functional. Vehicles at levels four and five on the other hand, have the ability to drive autonomously, only relying on their on-board systems to realize their driving capability. The difference between level four and five is, however, that in level four a driver still has to be present in the vehicle because the vehicle, although able to drive without human intervention, might not be able to do so in all driving modes (i.e. it is capable to drive on its own in a highway scenario but requiring human support in complicated urban settings). In level five we then find vehicles capable of what might be called ”true full automation”, meaning that a vehicle, via its on-board systems, is able to perform all driving tasks on its own and no human driver is needed anymore [2]. As examples from level five one could imagine a taxi service for non-drivers or a car-share repositioning system. Our focus for this paper lies on the last two levels, four and five, which represents cars driving autonomously by themselves.

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
Human driver **monitors the driving environment**						
0	No Automation	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
Automated driving system **("system") monitors the driving environment**						
3	Conditional Automation	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	High Automation	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	Full Automation	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Figure 1.1: Table of automation classification by the J3016 SAE standard. Source: [2]

An automated vehicle of level four and five, can be described into two different types. The first type is „autonomous“, meaning the car relies only on on-board sensory equipment to carry out their automated functions. The second type is a cooperative car, which is able to communicate. This communication occurs either between vehicles (vehicle to vehicle, V2V) or between the vehicle and roadside infrastructure (vehicle to infrastructure, V2I) [3]. This allows to use optimised routing algorithms and other means of traffic flow controls. In order to examine the benefits from an adoption of self-driving cars, studies focus mainly on cooperative cars.

Car companies like Tesla or Volvo as well as tech-companies like Google or Uber are taking steps into the direction of fully autonomous vehicle systems. The promises are high regarding all sorts of benefits of these self-driving cars. A study of the Columbia University for example is providing a glimpse of the potential effects. It suggested that a fleet of just 9000 autonomous cars could replace every taxi in New York City with an average waiting time of 36 seconds and an approximate cost of $0.50 per mile [4]. To get a better understanding of possible economic outcomes of such transportation systems, this paper examines the following question:

„What economical benefits does the proliferation of fully self-driving cars yield and what challenges are to be expected?“

In order to answer this question, possible economic effects of the availability and performance of self-driving cars have to be identified. In a first step, we look at those effects and their perceived benefits in isolation. However, to come to a full picture of economic impact, those effects have to be discussed in connection with possible scenarios of adoption of self-driving cars. The evaluation is based on four different scenarios of adoption found in the literature [5]. This evaluation then forms the basis for further conclusions.

This paper focuses on the economical aspects of self-driving cars. For this reason the methodology of how economic impacts are measured and quantified has to be defined in a first step (Section 1.2). It follows a examination of those economical impacts and a presentation of the findings (Section 1.3). After that, four scenarios of adoption are introduced (Section 1.4) to evaluate the previously examined economical impacts in the context of those four scenarios (Section 1.5). The paper ends with an outlook in terms of challenges and risks for future developments (Section 7.2.5) and a final summary and conclusion (Section 1.7).

1.2 Measuring economic impact

This section gives an overview of approaches used to measure the economic impact of a certain technology. It discusses the core concepts and challenges of the respective practices.
There are two types of approaches to measure economic impact. The first type of approaches focuses on measuring this impact in monetary terms. This is the realm of the traditional cost-benefit analysis, where the total costs of a new technology are compared to the total benefits one expects the new technology to have. The total expected benefits are then often contrasted with the current status quo in fields where the new technology is applicable. Such an approach of course necessitates that the consequences the introduction of a technology might have, are suitable to be captured in terms of money. In the field of self-driving cars this mainly includes effects in the areas of fuel consumption [6], safety [7] and productivity [8] where one either deals directly with money as an indicator, when for example looking at fuel expenditures, or the monetary value of measured indicators are easy to derive (e.g. assigning a monetary value to time spent working in case of looking at productivity).
The second type of approaches on the other hand, is concerned with examining effects that are difficult to express in monetary terms. In case of self-driving cars, one prominent example are effects that influence the environment. Here it is very difficult to directly assign a monetary value to observed indicators, because it is not evident what the monetary consequences of a certain effect might be. For example, one may know that it is considered a good thing to produce less greenhouse gas emissions, but how this might translate to monetary benefits, or costs in case of not doing so, is not clear. Also into this category fall effects that have an impact on the human psyche. It is for example quite hard to attribute a monetary value to stress that is possibly reduced by not having to drive your car by yourself. Nevertheless such effects exist and and are relevant thereby necessitating examination.
What people typically do to analyse such non-monetary effects is actually quite similar to a cost-benefit analysis in the classical sense, just without an explicitly monetary indicator. In case of greenhouse gas emissions, one looks at the expected benefits and costs not in monetary terms but uses empirically measurable indicators (e.g. amounts of pollutant emitted per mile driven) to assess the performance of a the new technology in comparison to the baseline scenario of the status quo. For psychological effects such as stress, techniques from the social sciences such as questionnaires or interviews are used to measure the relevant indicators. With self-driving cars such an approach is primarily followed to assess their impact on the environment [9] as well as on traffic performance [10].
Finally there is to say that, because self-driving cars are not in operation in significant numbers today, most studies use simulations paired with estimation and statistics to gather their data and perform their analysis.

1.3 Economic Impacts

There are numerous areas where the proliferation of self-driving cars might lead to benefits for individuals as well as society as a whole. The academic work in this field is broad and diverse, covering a wide range of possible effects and producing a large number of publications every year. This is due to the fact that vehicle automation technologies receive a lot of attention from the public in general, as well as the speed by which the technological progress in the field is realized. Nevertheless, for our paper, we had to restrict the effects covered to a sensible amount, as not to go beyond its scope. To obtain

a most current view on the findings in the literature for the effects we are interested in, we decided to focus on scientific papers being published during the last three years.
Two types of effects are examined. First we have effects that directly stem from the new and different performance features that self-driving cars offer in comparison to conventional cars. Here we cover five areas in which the expected effects are thought to be most relevant. Those are the impact of self-driving cars on traffic safety, energy consumption, traffic performance, individual productivity, and environment.
The second kind of effects capture impacts that are not directly related to the performance features of self-driving cars, but rather result from the availability of self-driving cars to human users. We examine two effects that fall into this category. The first effect is an expanded mobility for non-drivers. Effects in this domain are based on the assumption that cars that do not need a human driver will enhance the possibilities for mobility of people not able to drive themselves, be it because of illness or age. The second kind of indirect effects are so called macro-economic ones. The assumption here is that the proliferation of self-driving cars will have a profound influence on the structural organisation of an economy. Self-driving cars are assumed to influence the job market (by making human drivers obsolete), the insurance sector (by altering the needs for car insurances) as well as the car manufacturing sector (by creating new demand and usage patterns for cars in general).
In the following subsections we briefly review the most important findings for each effect during this time period.

1.3.1 Impact on safety

Self-driving cars are expected to bring major benefits in the area of traffic safety. The main arguments here is that eliminating the human driver from the equation and replacing it with an automated system, will severely reduce the amount of accidents caused by human error [11]. A blue paper from Morgan Stanley [8] estimates that, based on data from the US traffic authorities, about 90% of car accidents are caused by human error. To gauge the potential savings from self-driving cars they make a simple argument, namely that if 90% of car crashes are caused by human error, not having a human driver at all will reduce the number of accidents (and associated costs) by that amount. In monetary terms, the total yearly costs generated by car crashes in the U.S. is reported to be $625 billion. By applying their simple logic, Shanker et al. estimate the amount of possible savings to be around $563 billion per year (90% of those $625 billion). This potential savings are a rather optimistic estimate, based on the assumption that automation technology is capable of preventing all accidents caused by human error. To contrast this finding, the Casualty Actuarial Society's Automated Vehicle Task Force [12] revisited the results from the Morgan Stanley study and found that about 49% of the accidents attributed to human error were due to conditions that an automated system might also have trouble to handle (e.g. inclement weather, malfunctioning equipment). Applying that finding to Morgan Stanley's estimate one arrives at a more realistic savings potential of about $306 billion, which is still a lot of money.
The Morgan Stanley's estimate assumes a 100% market penetration of self-driving cars. In scientific literature, one also finds studies that deal with diverse market share ratios. Fagnant & Kockelman for example estimate the potential benefits of self driving cars to traffic safety for a market penetration of 10% , 50% as well as 90%. Their results suggest that potential savings increase with a higher market penetration of self-driving vehicles. For 10% , 50%, and 90% market share they arrive at comprehensive cost savings from crash reduction of about $ 17.7 billion, $ 158.1 billion, and $ 355.6 billion respectively, which is an amount roughly in the magnitude of the Casualty Actuarial Society' s estimate.

1.3.2 Impact on energy consumption

The main impact of self-driving car technology on energy consumption, a reduction of fuel consumption, are expected to be realized due to heightened fuel efficiency through automated eco-driving, platooning, de-emphasis of vehicle performance, and a possible reduction of vehicle-size [13]. Additionally, people consider the effects of smoother traffic flows and better routing possibilities, brought about by the capabilities of connected self-driving cars [15]. By applying drive-cycle simulation models, the authors estimate the total reduction in fuel consumption to be in the range of roughly 33% compared to conventional cars.
In terms of money, the Morgan Stanley blue paper [8] estimates that for the U.S., such a reduction of about 30% in fuel consumption would result in potential yearly savings of about $158 billion, assuming a 30% reduction of the yearly fuel costs of $535 billion.
What is not covered by these findings are the potential impacts that the use of non-fossil fuels (i.e. electricity) might have when combined with a vehicle automation technology. Such impacts are beyond the scope of this paper.

1.3.3 Impact on traffic performance

Main benefits for traffic performance gained by self-driving vehicles are expected to result from smaller vehicle-to-vehicle gaps as well as an overall smoother traffic flow [10]. To fully reap those benefits, it is necessary that the vehicles are acting in a cooperative way, exchanging information on their current speed and overall traffic conditions. Assuming a 100% self-driving vehicle scenario for highways in their microscopic simulation model, Aria, Olstam & Schwietering show that potential benefits amount to 8% increased average vehicle density, 8.5% increased average speed and 9% reduced travel time for any given road section, resulting in an overall improved traffic performance.
Such an improved performance might lead to monetary savings for fuel consumption through congestion reduction. The Morgan Stanley paper [8] estimates, that for the U.S., a yearly amount of $11 billion might be saved due to increased traffic performance resulting in less congested roads and therefore less time spent in congestion in general.

1.3.4 Impact on productivity

For benefits to individual productivity stemming from the proliferation of self-driving cars, the argument is quite easy to follow. If in the future, we do not have to drive our cars by ourselves, the time spent in our cars formerly used to drive can be used for other, possibly productive, endeavors. Car manufacturers imagine the interior of self-driving cars to serve as mobile living rooms as well as offices, permitting people to work while they travel from one location to another.
In terms of money, Morgan Stanley [8] estimates, based on statistical data from U.S. traffic authorities, that the potential monetary value of time freed up by not having to drive, and that therefore can be spent working, amounts to about $422 billion per year. They arrive at this number by applying a monetary value of $25 to each hour of work and assume that around 90% of the 18.8 billion hours spent in vehicles yearly will be used for productive work.

1.3.5 Impact on the environment

The main hopes for a beneficial impact of self-driving cars on the environment result from less pollutant emissions due to an improved fuel efficiency. This improved efficiency is realized through automated eco-driving, reduced vehicle sizes and performance requirements,

as well as platooning effects [13]. Further potential benefits arise in the case of shared automated vehicles, where it is envisioned that less vehicles are able to transport more people, thereby reducing the amount of vehicle miles travelled and vehicles in operation [6].
As for results, Fagnant & Kockelman estimate, using agent-based simulation models, that the introduction of shared automated vehicles, accounting for vehicle operation as well as manufacturing, might bring significant benefits in terms of pollutant emissions. Compared to an ordinary car, the shared automated vehicle used in their simulations would reduce the average emissions of pollutants by about 22% (including CO_2 and NO_x among others). As a caveat, they mention that the availability of shared automated vehicles might lead to an increased travel demand that possibly counteracts the potential benefits by generating a higher traffic volume overall. Additionally, shared automated vehicles might introduce about 11% more vehicle miles travelled due to relocating empty vehicles for other trips. In general, these results are difficult to translate to monetary terms, as the potential future costs of emitting pollutants into the environment in terms of climate change or medical consequences are controversially discussed and not really clear. Consequently, we did not find any references that estimate the monetary benefits of reducing pollutant emissions.
Again we do not cover the impact of using electricity as the main energy source to fuel automated vehicles, which in combination might further enhance the ecological benefits of self-driving cars.

1.3.6 Impact on mobility for non-drivers

The proliferation of self-driving cars is expected to bring about new possibilities for the mobility of non-drivers. As suddenly one does not need a human driver anymore, this opens up individual mobility for persons that are not capable to drive themselves. Possible causes for such an inability might be age (young / elderly people) or driving-restricting medical conditions (disabilities like blindness or epilepsy). A self-driving car provides people in those groups a means to travel without having to drive themselves. The reasoning behind this development is rather trivial. As people are presented with new possibilities, it is plausible to assume that they will want to use those new possibilities. This might increase the overall demand for travel and road transportation [14].
In their study for the expected travel demand for U.S. citizens previously unable to drive, Harper et al. assume three demand groups (called demand wedges in their paper) to estimate the additional travel demand that might be created. Group number one assumes that non-drivers (i.e. people without a driving license) will travel about the same amount as people that are currently able to drive. Group number two assumes that elderly people without travel-restricting medical conditions will travel roughly the same amount as younger people, while demand group number three expects people with driving-restricting medical conditions to travel about the same amount as people not suffering from those conditions. Starting from these assumptions, using statistical data from the U.S. National Household Transportation survey, and assuming a 100% automated vehicle scenario they predict that the ubiquitous availability of self-driving cars might result in an additional 295 billion miles traveled per year for the U.S. population. This represents a 14% increase in yearly vehicle miles traveled.

1.3.7 Macroeconomic effects

Under macroeconomic effects we understand effects that do not influence individual areas of the transportation system alone, but shape the structural organization of an economy as a whole. As such, these effects are difficult to measure exactly, since numerous assumptions about the future are required. The arguments in this subsection are not meant to be

comprehensive. We look at three areas where we think the proliferation of self-driving cars might have a profound impact. Those areas are the job market for human drivers, car-manufacturing industry, and car-insurance sector.

1.3.7.1 Job market for human drivers

For the job market for human drivers, severe consequences are to be expected should self-driving cars be available in great numbers. As those cars don' t need a human operator anymore, as a consequence, a lot of people might lose their jobs. Looking at the U.S. in the year 2014, the U.S. bureau of labor statistics reports 233' 700 people employed as either taxi drivers or chauffeurs [16] and 1' 797' 700 people working as truck drivers [17]. If self-driving vehicles are able to replace all those people, we look at about 2 million additional jobless individuals.

1.3.7.2 Car-manufacturing industry

The car manufacturing industry is also going to be facing great challenges if self-driving cars become more common. PricewaterhouseCoopers estimates, that when the adoption of self-driving cars nears 100%, the amount of vehicles on the roads might be reduced by up to 99% [18]. This will mean that car-manufacturers will either have to downsize their business dramatically, possibly resulting in less jobs, or come up with new products and business models to compensate for the expected reduction in car sales.
Not only will car-manufacturers be experiencing problems due to lower sales, they will most likely also face more competition in the market because new players are likely to enter the scene. With tech-giants, like Google or Apple, planning to release their own self-driving vehicle solutions [19] in the near future, traditional car-manufacturers will either have to adapt to the new realities of the car market or face the prospect of going out of business.

1.3.7.3 Car-insurance sector

Another profound effect is likely to be felt by the car-insurance sector. A KPMG study [20], by conducting a survey with senior insurance executives, estimates that during the next 25 years the market for car-insurance is likely to shrink to less than 40% of its current size. This effect is expected due to a decline in insurance premiums for car accidents, first because self-driving cars are likely to be safer that currently available ones, and secondly because the overall amount of cars on the road is likely to be reduced. The study also notes that we are likely to observe a shift in insurance patterns, away from insurance against human failures to insurance of product liability and technical failures.
As for the car-manufacturing industry, these aforementioned developments might force insurance-companies to downsize their business and reduce employment. At the same time, a changing market also means new possibilities for innovative business models to be established. Those emergent business models might attract new players to the insurance market, thereby increasing competition even further.

1.4 Scenarios of adoption

Four scenarios of adoption for self-driving cars are discussed in the following subsections. The adoption of autonomous vehicles will affect everyday life of human beings and their usage of mobility. In which ways exactly the introduction of such technology will alter the way humans think about, and enact, their new mobility is less clear. To capture possible avenues of development, Brian A. Johnson of Barclays defines four scenarios of

adoption for self-driving cars [5]. Those scenarios reflect different mobility patterns, and their effect on car usage, that might arise by the proliferation of self-driving cars. As these scenarios differ in the extent by which self-driving cars are going to alter mobility and travel demand, possible beneficial economic effects may result from one scenario, but not from another.

1.4.1 Traditional Vehicles

This scenario serves as a baseline in comparison to the other three scenarios. It assumes that in case of a proliferation of self-driving cars, people' s mobility behavior will likely stay the way it is today. The same is assumed for car ownership, namely that cars are privately owned assets. Since people are still going to keep their privately owned vehicle and will not change their travelling behavior, there is no change in the amount of cars on the street compared to today. The distance every car is travelling per year stays the same as well (cf. figure 1.2).

1.4.2 Family Autonomous Vehicles (FAV)

The family car is a concept, where not every family member has to own his/her own car, rather the family owns one car that is at every family member's disposal. An example use might look like this: In the morning, the father is using the car to drive to work. Once he arrives at work, he sends the car back home, where the children might use it to get to school or other social activities (cf. figure 1.2). Since not every family member has to have their own car, the average number of cars per household drops from 2.4 to 1.2 units. This leads to less cars on street, but doubles the amount of miles per car due to empty trips necessary to relocate cars between different users.

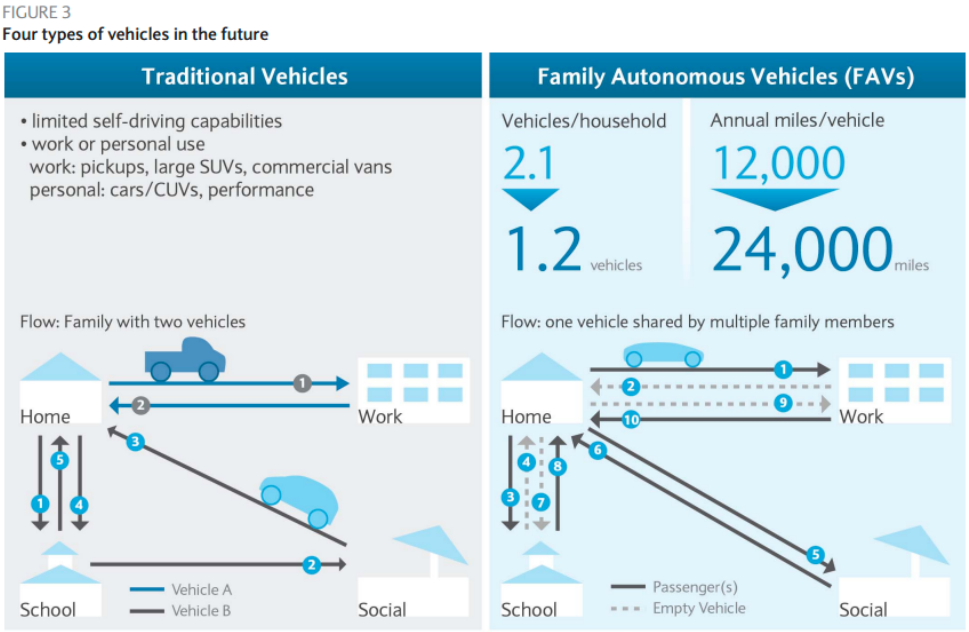


Figure 1.2: Traditional and Family Shared Vehicles. Source:[5]

1.4.3 Shared Autonomous Vehicles (SAV)

In the scenario of shared autonomous vehicles, the car is not owned by a private person anymore. It is either publicly owned or owned by a company. The autonomous vehicles are available on demand and are routed to pick you up, when the need arises. It is imagined that self-driving cars in this scenario serve as some sort of „robotaxi", autonomously

transporting people where they need to go. Uber is researching in this field, since expectations are high, as shown in the Columbia University study (2013). As shown in figure 1.3, one shared vehicle is accumulating around 5 times more miles in a year than a conventional vehicle, but it is, in contrast, replacing 9 privately owned vehicles.

1.4.4 Pooled Shared Autonomous Vehicles (PSAV)

The fourth scenario is about pooled shared autonomous vehicles. It is similar to the SAV scenario, but, in contrast, a user of such a vehicle is not using it alone by herself. This can be seen as some sort of public transportation, but, however, it' s not clear if the vehicles are publicly owned or owned by a company. The concept is that a vehicle picks up different people along a given trip, meaning one person is sharing its ride with other people. The car will drop off or pick up new people that share a similar route (as shown in figure 1.3). This idea of vehicle sharing results in a so-called perpetual ride. From a cars' s perspective this is a ride that never ends and, ideally, is never empty. Such a system leads to a very efficient way of transportation. One PSAV is replacing about 15-18 conventional vehicles, but the annual accumulated miles per vehicle are the same as in the SAV scenario.

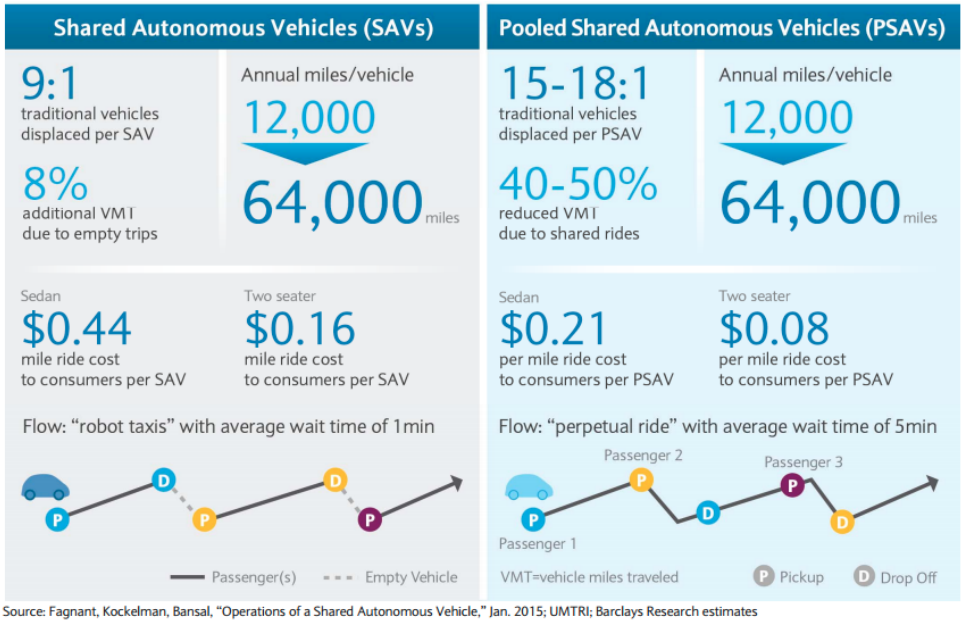


Figure 1.3: Shared autonomous vehicles and pooled shared autonomous vehicles. Source: [5, 21]

1.5 Evaluation and Discussion

By combining the four scenarios of adoption with the previously introduced economic impacts, four different evaluations in regard of economical effects can be derived. The traditional vehicle is used as a baseline in this evaluation, which the other scenarios are going to be compared against. However, some effects are more generic and are relevant in all or more than one scenario, due to inherent features that self-driving cars are offering. A brief overview of the comparison of the scenarios is shown in table 1.1. The discussion of each scenario are provided in the following subsections.

Impacts:	Scenarios of Adoption for Autonomous Vehicles: **Traditional**	**Family**	**Shared**	**Pooled Shared**
Safety	+	+	+	+
Fuel consumption/mile	+	+	+	+
Mobility for non-drivers	+	+	+	+
Job-Market	-	-	-	-
Insurance sector	+/-	+/-	+/-	+/-
Traffic performance	+	-	+	++
Energy consumption	+	-	+	++
Environmental impact	+	+/-	+	++
Impact on car sales	-	-	–	–
Comfort	++	++	-	–
Productivity	++	++	-	–

Table 1.1: Evaluation of the impacts regarding the four scenarios of adoption.

1.5.1 Evaluation of Traditional Vehicles

The first scenario, traditional vehicles, acts as a baseline for the other scenarios. Effects found for this scenario can be considered to apply to the other scenarios as well, since they are based on the technological progress of autonomous vehicles. There are a few positive effects that can be observed for traffic safety, fuel economy per mile, productivity, and mobility for non-drivers. Traffic safety is going to increase, due to the fact that the number of crashes resulting from human error are going to drop and, therefore, less costs are generated due to crashes. Fuel economy per mile is going to be much more efficient, since vehicles use eco-drive-mode and save fuel by platooning. Since drivers are not drivers in the conventional sense, more like passengers, they are able to perform working task while travelling from location to location. This leads to an increase in overall productivity. Additional mobility for non-drivers can be considered as a positive economic effect as well, since a previously untapped market segment is now accessible for companies. However, there are a few drawbacks and uncertainties. One aspect is that, since non-drivers are enabled to use automated vehicles on their own, travelling demand in those groups will increase, which results in a negative side effect for the environment due to more miles travelled overall. Another negative effect can be seen for the job market. Since no human driver is needed in a taxi or other transportation vehicles anymore, this might leads to a higher unemployment rate as a short term effect. There is no guarantee that re-education is a valid and feasible option for every person affected by the shifting demands of the job market. Another effect which is related to traffic safety, and hard to predict, is the effect on the insurance sector. Are existing insurance models still valid and used, if there are less or no car crashes in the future? Or are there going to be other models which are replacing current models? However, it is safe to say that insurance models are going to be impacted in one way or another, especially in the long-term, when no human driver is needed on the road anymore and accidents are claimed to be a relict of the past.

1.5.2 Evaluation of the Autonomous Family Vehicle

In this scenario, no additional advantages were observed compared to the baseline. However, there are a few possible extra disadvantages and uncertainties. Since the family is sharing a car, the volume of cars in use is declining about roughly 40%, but the remaining cars in use are going to double their annual travelling distance, due to empty rides. Comparing both of the effects, we see a decrease of 40% vehicles in use against a 100% increase in annual miles per vehicle. From this, it is obvious that there are going to be

more miles travelled a year and therefore, energy consumption is going to increase. There will be more vehicles on the road as well, even though there are less vehicles in use. This fact is affecting the traffic performance in a negative way as well. The demand for new vehicles is going to decrease since less vehicles are owned. This will affect the whole car manufacturing economy and their labour in the end, which could lead to additional job losses in the car industries on top of the job losses due to a decreased demand for human drivers. The environmental impact is hard to estimate. There are positive impacts due to less vehicles in use and therefore less vehicles built. In contrast, more miles travelled lead to higher pollutant emissions.

1.5.3 Evaluation of the Shared Autonomous Vehicle

In the third scenario, using shared vehicles, the positive effects outweigh the negative ones. This is due to the fact that the ratio between vehicles in operation and annual miles traveled per vehicle is much better than in the baseline scenario. The replacement of conventional vehicles by a shared autonomous ones shows a ratio of 9:1 (i.e. one shared autonomous vehicle is replacing 9 conventional vehicles), whereas the increase in annual miles traveled per vehicle has a ratio of 5.33:1. This leads to less miles traveled annually in total, which has a positive effect on energy consumption. Other aspects that are influenced in a positive way, resulting from those better ratios, are traffic performance, due to less traffic in total, and environmental impact, due to less vehicles built and less pollutant emissions due to reduced energy consumption. An economic sector which is going to be influenced in a negative way is again the car manufacturing industry, since there will be a decline in the demand for new cars. Additionally, traditional car companies are challenged by new players in the mobility market, such as Google or Tesla.

1.5.4 Evaluation of the Pooled Shared Autonomous Vehicle

The scenario dealing with pooled shared vehicles can be interpreted as an alternative version of the shared autonomous vehicle scenario. It exhibits the same positive effects, but with an even better ratio of vehicles replaced. The number of vehicles in operation is decreasing even further, replacing 15-18 conventional vehicles by one pooled shared one. On the other hand, the annual traveled miles per vehicle stay the same as in the shared autonomous vehicle scenario. A pooled shared autonomous vehicle will again travel 5.33 times more miles per year than a conventional one. Therefore, energy consumption is reduced even more, due to less annual miles traveled in total. Traffic performance is better as well, due to the same fact combined with less vehicles in operation overall. The effect on the environment is going to be positive as well, since the demand for new vehicles decreases and therefore less vehicles are built. Additionally, pollutant emissions are reduced, due to less traffic. On the negative side, the car manufacturing industry is going to be impacted in an even more extreme way, due to a further decrease in demand for cars and again a further impact on the employment rate in this sector. Sharing a vehicle with other people might not be the most eligible thing for people. This results in reduced comfort while traveling, similarly to using public transport. This also means that using the vehicle as a working space is harder and therefore, the promised productivity increase might not be fully realized.

1.6 Challenges and Risks

Apart from the different economic effects self-driving cars might have, there are numerous other challenges that have to be faced in order to establish them as an accepted means of personal transportation.
First of all, there are people that do not see the need to drive as a necessary evil, but love to do it. Oldtimer and sports car enthusiasts will most likely not just stop driving around in their what might be called „legacy vehicles“[22]. Therefore we need a way to ensure, that conventional vehicles as well as self-driving ones will be able to coexist on our roads. Not only technical or legal concerns have to be considered, one has also to pay attention to the the perception of automated vehicles by drivers of conventional vehicles. The design of automated vehicles is in regard of system performance, as a result driving patterns, such as driving in the center of the lane, occur. Whereas, the driving behavior of human drivers is a results of intuition and other, not necessarily optimal, reasoning. Therefore, in a mixed traffic situation, two safety issues can occur: the human driver either misses important clues of automated vehicle on its intentions, or he expects the autonomous car to drive similar to a human-operated one [23]. A study provided initial evidence that safety issues can arise when unequipped vehicles' drivers encounter assisted driving behaviour [24].
Another challenge to be faced is to guarantee the security of the new self-driving cars from hackers and other parties of malicious intent [25]. Tightly coupled to these issues of security is the question of trust. A lot of people have trust issues regarding automated systems [26] and therefore regarding self-driving cars as well [27]. With this problem, the question of how to demonstrate the safety and reliability of self-driving car systems arises. In order to statistically demonstrate safety and reliability in terms of injuries and fatalities, the autonomous vehicles would have to travel for millions of miles. To alleviate these two concerns in society, self-driving car developers will have to come up with new innovative methods for demonstrating safety [28].

1.7 Summary and Conclusion

In our paper we explored the possible economic benefits the proliferation of self-driving cars might have for individuals as well as society as a whole. We started by introducing the concept of automated vehicle systems and refined it to the point of self-driving cars, which are the focus of this paper. Furthermore, we shortly explained two different approaches how economic benefits are measured and analysed in scientific literature. From this point on, we conducted a survey on recent scientific literature and derived the most important findings in those areas that we think are the ones where the most important benefits from the proliferation of self-driving cars might be realized. We focused on two types of effects, direct effects resulting from the additional performance features of self-driving cars as well as indirect effects, stemming from the availability of self-driving cars to a broader public. For direct effects, we considered the impact on traffic safety, energy consumption, traffic performance, individual productivity as well as the environment. In the domain of indirect effects, we looked at increased mobility for non-drivers as well as a selection of possible macroeconomic effects. We then explained four scenarios of adoption for vehicle automation technology taken found in the literature. Applying our findings to these scenarios, we evaluated the potential economic benefits that might be brought about by realising those scenarios.
As for conclusions, our first one is quite obvious. Self-driving cars will become a reality, possibly in the near future. Recent announcements by car-manufacturer Telsa, that they

plan to have a fully self-driving car available for purchase by the end of 2017 [29], provide a clear hint that we might find such cars on our roads rather soon.

Looking at the economic effects, that there are a lot of potential benefits to be gained by the proliferation of self-driving cars. Most notably is the increased safety for road traffic that is expected to be realized by getting rid of human drivers and their notoriously bad habits of being sleepy, distracted or drunk while driving. But not only will self-driving cars be safer than conventional ones, they will likely also use less fuel, provide additional time to work and bring us from point a to b in less time and possibly more comfort. Even better, self-driving cars might also be good for the environment and leave a smaller ecological footprint than conventional ones.

These prospects might sound fabulous, but there is always a downside. While the economical benefits are possibly large when comparing the performance of conventional and self-driving cars on an individual basis, the picture changes when looking at broader implications. Availability of self-driving cars will bring a new kind of individual mobility to people that were formerly not able to drive themselves. While this on its own is a desirable feature, such an improved mobility might lead to higher traffic volumes overall that possible negate the benefits (especially ecological ones) brought about by self-driving cars. Additionally, the availability of such cars, through macroeconomic effects, might severely impact the structural organisation of an economy as a whole. By making human drivers obsolete, a lot of people occupied in the driving profession might lose their jobs. Even more people might be impacted by the changes that are likely to affect the car-manufacturing industry as well as the insurance sector.

As always, such drastic changes also harbor chances. The proliferation of self-driving cars might give rise to numerous new business models in the field of individual and public transportation and beyond. These new business might possibly require new skills and expertise compared to those prevalent in today's world. Thereby it is not implausible to assume that, while some jobs might disappear, some other, previously unimagined, jobs and professions might emerge.

Regarding the likelihood of the four different scenarios of adoption for self-driving cars, we are not able to predict a clear winner. This is due to the reasoning that the adoption of such a new technology is largely shaped by the perceptions and attitudes that we humans have towards it. As with all human behavior, it is difficult to predict and might be influenced by factors that we are not even able to think of right now. It is possible that self-driving cars will replace the traditional cars we have today in a way that does not alter our preferences and behavior regarding transportation. On the other hand, it is equally likely that self-driving cars fundamentally transform our concepts of mobility and transportation as a whole. Nevertheless, one can state that the price for which self-driving cars will be available for purchase, or use, is going to play a major role. Whether these prices have to be paid to own such a new car yourself, or to one of the providers of new mobility solutions to use their cars, is largely dependent on the future business models that will establish themselves as dominant in the realm of self-driving cars and mobility.

This does not mean that we do not have a preference for one of the scenarios presented. In our point of view, the scenario where self-driving cars are used as pooled shared automated vehicles (scenario 4) is the most desirable one, because it leads to a more sustainable use of available resources. Not only will we need fewer cars overall, we will also use those cars in a more efficient way. Unfortunately, this scenario is also the one we expect to face the most resistance. This is due to the fact, that this scenario presents the most profound changes to how mobility is imagined. Not only will people not own their cars anymore, they will also have to share their rides with potentially unknown traveling companions. Here one can imagine that ingenious business people will create business models where such a private ride, while not owned privately, but at least used exclusively, will be available for a certain fee.

This leads us to our final conclusion: We imagine that no single scenario of adoption will be realised in isolation. More likely, we will see a mixture of business models catering to all four scenarios. There might be businesses specialising in shared rides, while car-manufacturers will still continue to sell cars to private owners. This is especially true for different traffic contexts. While it may make sense to have pooled, shared rides in urban centres and densely populated suburbs, in rural areas, those business models might not be applicable and people will need their own personal rides to go about their business.
Overall, the proliferation of self-driving cars harbors a large potential for economic benefits for individuals and society as a whole. Whether these benefits will be reaped to the full extent is largely dependent on how we, as humans, are able to embrace the chances such a new technology brings, while at the same time succeeding to absorb and counteract possible drawbacks.

Bibliography

[1] M. Brown. Autonomous Tractor Is Outstanding In Its Field. *Autonomous Tractor Is Outstanding In Its Field*, 2011, `https://www.wired.com/2011/09/autonomous-tractor-is-outstanding-in-its-field/`.

[2] Society of automotive engineers. Automated Driving. Levels of driving automation J3016 , 2014, `http://www.sae.org/misc/pdfs/automated_driving.pdf`.

[3] Jim Barbaresso, Cordahi Gustave, Dominie Garcia, Christopher Hill, Alex Jendzejec, and Karissa Wright. Intelligent Transportation Systems (ITS) Strategic Plan 2015-2019. Technical report, US Department of Transportation, 2014.

[4] Lawrence D. Burns, William C. Jordan, and Bonnie a. Scarborough. Transforming Personal Mobility. pages 1–43, 2013.

[5] Brian Johnson. Disruptive Mobility. *Barkley*, 2015.

[6] Daniel J. Fagnant and Kara Kockelman. Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167–181, 2015.

[7] Corey D. Harper, Chris T. Hendrickson, and Constantine Samaras. Cost and benefit estimates of partially-automated vehicle collision avoidance technologies. *Accident Analysis and Prevention*, 95:104–115, 2016.

[8] Ravi Shanker, Adam Jonas, Scott Devitt, Katy Huberty, Simon Flannery, William Greene, Benjamin Swinburne, Gregory Locraft, Adam Wood, Keith Weiss, Joseph Moore, Andrew Schenker, Paresh Jain, Yejay Ying, Shinji Kakiuchi, Ryosuke Hoshino, and Andrew Humphrey. Autonomous Cars: Self-Driving the New Auto Industry Paradigm. *Morgan Stanley Blue Paper*, pages 1–109, 2013.

[9] Daniel J. Fagnant and Kara M. Kockelman. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transportation Research Part C: Emerging Technologies*, 40:1–13, 2014.

[10] Erfan Aria, Johan Olstam, and Christoph Schwietering. Investigation of Automated Vehicle Effects on Driver's Behavior and Traffic Performance. *Transportation Research Procedia*, 15:761–770, 2016.

[11] Adriano Alessandrini, Andrea Campagna, Paolo Delle Site, Francesco Filippi, and Luca Persia. Automated Vehicles and the Rethinking of Mobility and Cities. *Transportation Research Procedia*, 5:145–160, 2015.

[12] Stienstra, M. B., Antol, M. L., Armstrong, S. D., Blair, G. C., Border, D. R., Bugbee, M. H., Liu, L. Casualty Actuarial Society E-Forum *The CAS*, Volume 1, 2014.

[13] Zia Wadud, Don MacKenzie, and Paul Leiby. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. *Transportation Research Part A: Policy and Practice*, 86:1–18, 2016.

[14] Corey D. Harper, Chris T. Hendrickson, Sonia Mangones, and Constantine Samaras. Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions. *Transportation Research Part C: Emerging Technologies*, 72:1–9, 2016.

[15] Chris Manzie, Harry Watson, and Saman Halgamuge. Fuel economy improvements for urban driving: Hybrid vs. intelligent vehicles. *Transportation Research Part C: Emerging Technologies*, 15(1):1–16, 2007.

[16] Bureau of Labor Statistics. US Bureau of Labor Statistics, Occupational Outlook Handbookm, Taxi Drivers, 2015, `http://www.bls.gov/ooh/transportation-and-material-moving/taxi-drivers-and-chauffeurs.htm`.

[17] Bureau of Labor Statistics. US Bureau of Labor Statistics, Occupational Outlook Handbook, Truck Drivers, 2015, `http://www.bls.gov/ooh/transportation-and-material-moving/heavy-and-tractor-trailer-truck-drivers.htm`.

[18] PricewaterhouseCoopers LLP. Autofacts, 2013.

[19] Computer World. Computerworld, Google: Autonomous cars coming 'relatively soon', 2016, `http://www.computerworld.com/article/3047514/car-tech/google-autonomous-cars-coming-relatively-soon.html`.

[20] KPMG. Automobile insurance in the era of autonomous vehicles. (October), 2015.

[21] Daniel Fagnant, Kara Kockelman, and Prateek Bansal. Operations of shared autonomous vehicle fleet for Austin, Texas market. *Transportation Research Record: Journal of the Transportation Research Board*, 2536:98–106, 2015.

[22] Daniel Howard. Public Perceptions of Self-driving Cars: The Case of Berkeley, California. *MS Transportation Engineering*, 2014(1):21, 2014.

[23] Roald J. van Loon and Marieke H. Martens. Automated Driving and its Effect on the Safety Ecosystem: How do Compatibility Issues Affect the Transition Period? *Procedia Manufacturing*, 3:3280–3285, 2015.

[24] Katharina Preuk, Eric Stemmler, Caroline Schießl, and Meike Jipp. Does assisted driving behavior lead to safety-critical encounters with unequipped vehicles' drivers? *Accident Analysis and Prevention*, 95:149–156, 2016.

[25] Society of automotive engineers. Major security challenges face makers and owners of connected cars, 2014, `http://articles.sae.org/13081/`.

[26] John D Lee and Katrina a See. Trust in automation: designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[27] Schaefer, K. E., & Scribner, D. R. Individual Differences, Trust, and Vehicle Autonomy: A pilot study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 786 790, 2015.

[28] Nidhi Kalra and Susan M. Paddock. Driving to Safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?. *RAND*, 2016.

[29] Aron Kessler. Elon Musk Says Self-Driving Tesla Cars Will Be in the U.S. by Summer. New York Times , 2015, http://www.nytimes.com/2015/03/20/business/elon-musk-says-self-driving-tesla-cars-will-be-in-the-us-by-summer.html?_r=0.

Chapter 2

Economics of Radio Towers

Bleyer Benedikt, Schneider Moritz, Triner Mirco

According to studies by Gartner [10] and Cisco [5] the market for mobile phones and the communication with mobile phones has increased as well as the usage patterns. Today mobile phones aren't used anymore only for calls or text messaging but also for watching movies over the internet. Forecasts predict that this way of usage will increase with future technologies like 5G. The network providers as well as the regulators have to face the challenges of this development and find approaches to satisfy the customer needs in the future. This report details three approaches to tackle those challenges. First, Mobile Network Operators (MNO) increase the number of cell towers (also known as base stations) in the mobile network infrastructure. Second, regulators change radiation levels emitted by the base stations. Third, MNOs integrate different mobile technologies like WiFi with 4G and 5G. Furthermore, different studies about the effects of radiation on humans are discussed and the costs and benefits of the approaches are compared with each other. In conclusion it is highly likely that not only one approach is needed or useful to engage the challenges, but a combination of the three proposed approaches.

Contents

2.1 Introduction

According to studies by Gartner [10] and Cisco [5] the market for mobile phones and the communication with mobile phones has increased. Those studies are also forecasting the growth will continue, at least in terms of the usage of the mobile networks.Today there are different technologies provided by the network providers which are then used in mobile or cellular networks, e.g. Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS) or Long Term Evolution (LTE).
Figure 2.1 shows the distribution of data traffic using the different mobile technologies is changing. According to a study by Cisco [5] the amount of data traffic used by older mobile technologies, 2G and 3G, is decreasing. While traffic of newer technologies, e.g. 4G or 5G, will increase over the next few years, due to the fact that today the mobile phone is also used for the consumption of videos in the Internet.

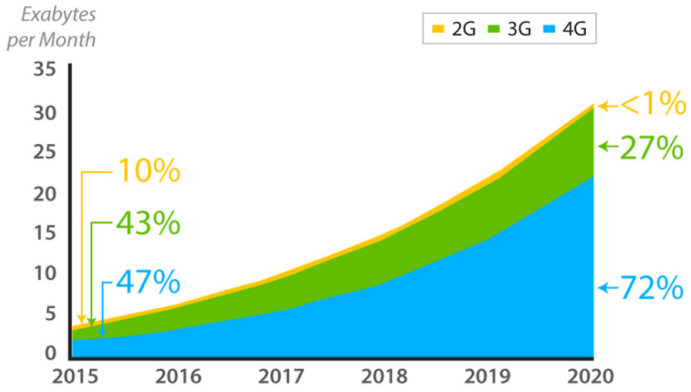


Figure 2.1: Mobile consumption 2015-2020 [5]

The mobile networks of former technologies weren't build to transfer that much data, nevertheless they are still heavily used. One example are tourists or business people, who don't own a mobile contract which allows them to use the mobile network to consume content of the internet via their mobile phone. Based on this assumption by the authors it will still be necessary to ensure the availability and quality of different mobile networks in parallel.
In many countries there is more than one mobile network, because different mobile network providers are in the market. Switzerland has three mobile network providers (Swisscom, Salt and Sunrise), which all maintain their own mobile network infrastructure. All this network providers had and have to face the challenges of popularity of the mobile phone. Over the last few years those different network providers have built and deployed many cell towers to provide and ensure good coverage and fast transfer rate, especially in the cities and their Central Business Districts (CBD). Figure 2.2 shows how many cell towers are installed in the CBD of Zurich. Important to notice is that at the moment every major mobile technology is using their own cell towers. Each circle stands for one cell towers (used as synonym for transmitter, base station and antenna). The color turquoise is used for cell towers of the technology GSM, pink for UMTS and blue for LTE (4G).
Figure 2.2 shows that there are already a lot of cell towers deployed in urban areas. Today the availability and quality of the mobile networks are very high. But to keep those characteristics on the same level over the next few years a continuous development of the network would be necessary according to the mobile network providers. All parties agree on the fact, that development and improvement of the mobile network is necessary to satisfy customer needs over the next few years. But in terms of methods or approaches to do that, there isn't a broad agreement across the different parties. Therefore the research questions of this report are the following:

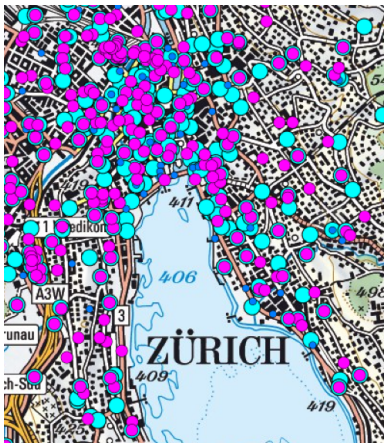


Figure 2.2: Antenna positions in Zurich [1]

1. What approaches are available to satisfy the customer needs in mobile communication?
2. What are the advantages and disadvantages of those approaches?
3. How are the effects of this approaches on the radiation level?
4. What are the effects of radiation on humans?

According to a literature review there are three main options possible to develop and improve the mobile networks. The first option is to build more cell towers. Another option would be to change the radiation regulations for existing cell towers to enhance their capabilities. The third option consists of the integration of the different mobile technologies, for instance WiFi.
To provide more details to those options along with details to answer the other questions this report is structured in the following way. Section 2.2 provides details of the evolution of mobile communication and the used network infrastructure and technologies. After explaining different cells types of a mobile networks in section 2.2.2, section 2.3 covers all the details about the proposed approaches, followed by the basic knowledge about radiation and the effect on human beings in section 2.4. A comparison of the costs and benefits for the approaches can be found in section 2.5. In section 9.5 the answers for the research questions are provided.

2.2 Structure and Function Cellular Networks

This section gives an overview of the mobile technology evolution, the functionality of a mobile network, provide a definition of the various components and how they are connected to each other.

2.2.1 Evolution of network technologies from 0G to 5G

In 1971 the Autoradiopuhelin launched the first mobile phone network in Finland as studies from Bhalla shows [2]. One year later the Germans launched their B-Netz for mobile calls. Those mobile telephones were of great size, therefore they were usually mounted into cars. Some models exist which were integrated into briefcases.
In the 1980s the first generation replaced the 0G technology according to Bhalla [2]. The first generation used analog radio signals and was modulated to 150MHz. With Frequency Division Multiple Access (FDMA) radio signals were sent between radio towers. The first generation had poor voice links and no security features.

Mobile phones of the second generation are generally smaller than those of the first generation. The second generation cellular networks were launched with GSM. Which made roaming around the globe possible. It uses Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA). GSM brought an increase of bandwidth, higher voice quality and less battery consumption, also Short Message Service (SMS) were possible over GSM [2]. GSM is split into three subsystems, Base Station Subsystem (BSS), Network Subsystem (NSS), Intelligent Network Subsystem (IN) [23].
The BSS includes a Mobile Station (MS), a Base Transceiver Station (BTS) and a Base station controller (BSC) according to Sauter [23]. The MS is connecting to the BTS. Therefor the BTS is equipped with all the necessary radio equipment to communicate with the MS. The BTS is forming the radio cell which can be between 100m and 35km. The BTS is controlled by the BSC. Radio frequencies and handover handling between BTS within the BSS has been done by the BSC.

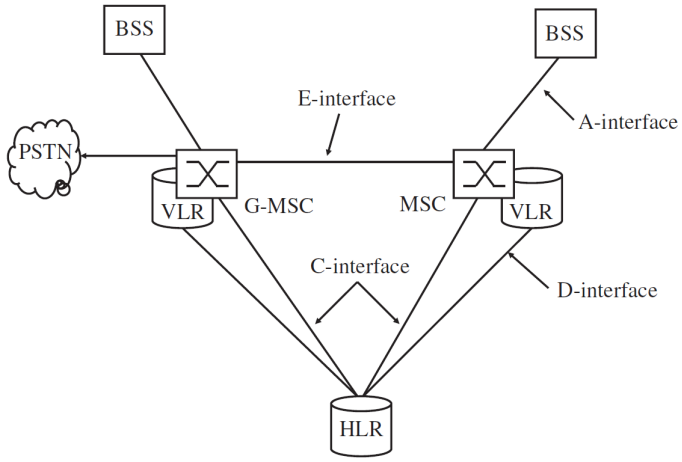


Figure 2.3: GSM Infrastructure [23]

Figure 2.3 shows a Network Subsystem which includes a Mobile Switching Center (MSC), a Home Location Register (HLR) and a Visitor Location Register (VLR) [23]. The MSC manages all the subscribers and the connections. It is connected to the BSS with the A-interface which is used for signaling and speech. The A-interface is normally a optical fiber, because the distance between an MSC and a BSS can be up to 100km. The MCSs are connected with the E-interface which is also used for signaling and speech. The C-interface connects the MSC and the HLR for signaling only. Each MSC has a VLR which has a record for every subscriber of the MSC. The HLR is the subscriber database of a GSM Network.
The Intelligent Network Subsystem (IN) provides more functions for billing [23]. One example is prepaid. Additionally, the IN allows to have different prices for phone calls, for different Regions or Providers.
With General Packet Radio Service (GPRS) the 2.5 generation was invented which added a packet domain switching to the circuit switch domain [2]. Also GPRS allows the billing for the transferred megabytes. With the past technologies the billing was made for the time used.
Figure 2.4 shows a GPRS Network which is providing additional Elements to the GSM Infrastructure. There is a Packet Control Unit (PCU) which is responsible for packet-switching and assigning time slots to subscribers [23]. The PCU is a counterpart of the BSC. The Serving GPRS Support Node (SGSN) is responsible for circuit-switching. The SGSN is a counterpart of an MSC. Gateway GPRS Support Node (GGSN) routes data between the radio access network and the core network.

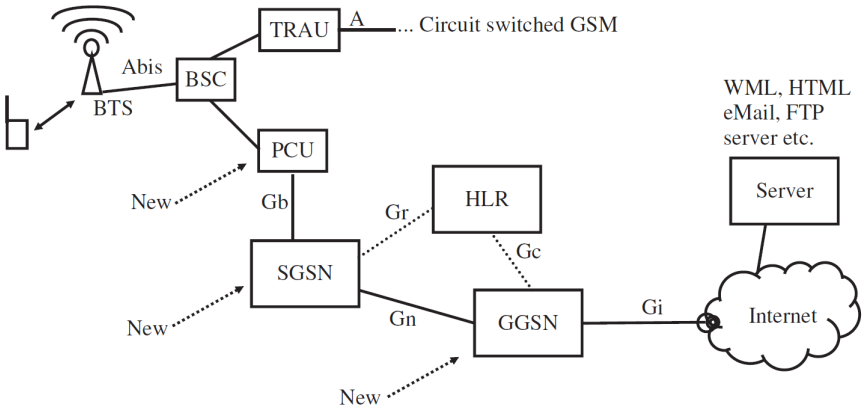


Figure 2.4: GPRS Infrastructure [23]

With the Third Generation additional services were available, like broadcast wireless data and video calls. The aim of 3G was more coverage and growth with a minimum investment. The UMTS which is used in 3G uses TDMA, CDMA and FDMA [2].

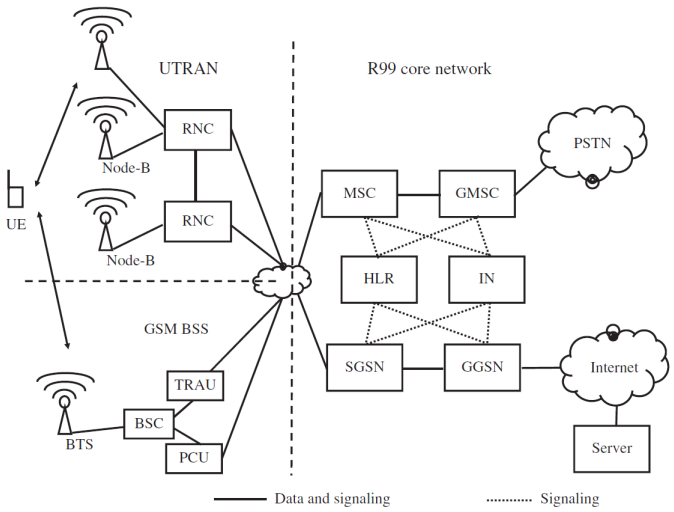


Figure 2.5: UMTS Infrastructure [23]

Figure 2.5 shows an UMTS infrastructure. The UMTS network implementation calls a User Equipment (UE). The UE is connecting to a Node-B which is equipped with antennas. The Heart of the UMTS Network is a Radio Network Controller (RNC). To an RNC several hundred Node-Bs can be connected. At the beginning of the UMTS deployment those connections were via fixed-line links with a bandwidth of 2Mbit/s. Today the links are changed to fiber or ethernet. [23]
The fourth generation is an extension of 3G with more services and bandwidth. It was launched in 2010 [2]. With 4G the high quality audio-video streaming over Internet Protocol became possible. Bandwidth speeds up to 1Gbps over fixed stations are possible. In the fourth generation there is a change in the access control according to Sauter [23]. Multiple Input Multiple Output (MIMO) is used which allows simultaneously connections to one device. The device is therefore equipped with multiple antennas. Figure shows a mobile device connecting to a eNode-B. The eNode-B's are doing the handover among one another without contacting a master unit. Also Quality of Service (QoS) is managed by the eNode-B's. A Mobility Management Entity (MME) is doing the Handover to other technologies like GSM. A Serving Gateway (Serving-GW) is responsible for managing user data tunnels between a Packet Data Network Gateway (PDN-GW) and the eNode-Bs in the radio network. The PDN-GW is the gateway router to the internet. Also

intranet connectivity for large companies encrypted data transformation the PDN-GW is responsible. IP-Addresses were provided to the mobile devices by the PDN-GW. A Home Subscriber Server (HSS) is a subscribers database and is referred to the HLR.

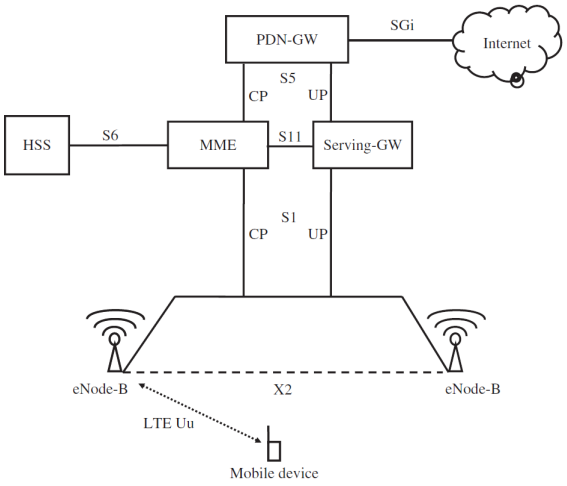


Figure 2.6: LTE Infrastructure [23]

5G will be presented in 2020 according to Osseiran [20]. In 2020, 50 billion connected devices have to be supported, which are then part of the Internet of Things (IoT). New services like e-learning and e-health will be heavily used. The connections have to provide far more stringent latency and reliability, a wide range of data rates, up to multiple gigabits per second, and a support of very long battery lifetimes. To achieve the integration of the new services and devices the project "'METIS"' was founded. They are using a bottom-up process for developing new radio concepts and for optimizing the support for future application needs. In a top-down approach, application's needs and relevant use cases were evaluated. "'METIS"' came up with some requirements. There are expected data rates from one to ten GB/s. There is an increase of subscribers to an Access Point up to 300'000. The latency must be less than 5 ms for safety reasons and the battery life time must be at least one decade.

2.2.2 Macrocellular and Femtocell Networks

With the technological progress the user demands have changed and the internet has become a mandatory feature. Therefore, the network providers had to change their deployment strategy. Before the change, they were focusing on wide area voice coverage. But after that, they had to increase the bandwidth for new needs as section 2.2 shows. Especially in hot-zones, enterprise building or homes the coverage do not meet the demands [16]. To cover all the different needs, the service providers can use different deployment strategies as Table 2.1 shows.

Table 2.1: Different types of cells [16]

Type of Cell	Typical range	Typical Environment	Backhaul
Macrocell	2 - 30 km	Outdoor, wide area	Operator controlled
Microcell	200 m - 2 km	Outdoor, wide area	Operator controlled
Piccocell	Up to 200 m	Indoor public locations	Mostly operator controlled
Femtocell/Homecell	Up to 20 m	Indoor residential of business premises	End user of third party controlled

Macrocell deployment is used to cover wide areas. Especially countrysides with few subscribers allotted in a large area the service providers can provide good coverage and speed with just one cell. If there are lot of mountains the number of cells has to be increased.
The micro cell deployment is like the macro cell deployment only for outdoor usage. Micro cells can cover areas from 200 m up to 2 km and are perfect to cover crowded places like railway stations [16].
To cover shopping centers and other hot zones the cellular operators can deploy piccocells. Those piccocells are co-channel with the macrocells, which means that they are using the same frequency [16]. For preventing interference, the piccocells are operated with small antenna power, which results in a coverage of about 200 m [16].
Femtocells or homecells are installed by the user. Figure 2.7 shows the cells connecting directly via the Internet to the home cell controller, which builds the entry point to the operators network. Those indoor cells have a coverage of 20m [6]. Only a closed subscriber group can connect to those cells. To avoid an increase of handovers, the femtocells are deployed without leakage outside the building. This is done by a reduction of the cell size.

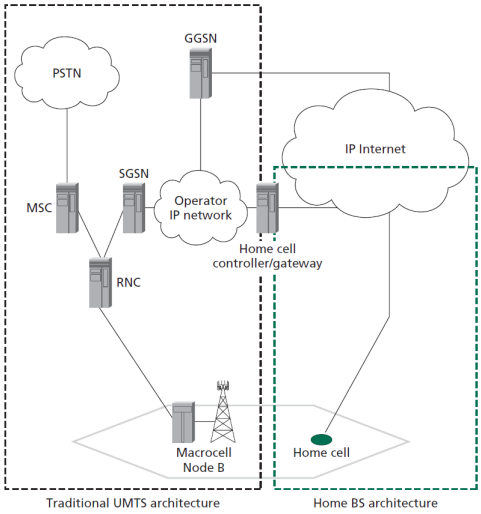


Figure 2.7: Home cell deployment [6]

2.3 Approaches

In the 1940s, Claude Shannon developed the concept of channel capacity and then formulated a complete theory of information and its transmission [27]. Shannon stated that if the data transmission has to be increased, either more antennas must be used, a larger bandwidth should be used, or the transmission power should be increased. 70 years later, Shannon's law is still valid for mobile providers. The bandwidth of mobile networks is limited since it is based on the spectral efficiency. The spectral efficiency is constantly increased by new mobile radio technologies. As a result, more transmission capacity can be provided with the same transmission power of a base station. However, this improvement can not fully compensate the strongly increasing data traffic in the mobile networks (currently an annual doubling). LTE-Advanced (a further development of LTE) is expected to be a factor of three times more efficient than HSPA+. With the current growth of data traffic by a factor of two times per year, this technical improvement is only sufficient to compensate for the traffic increase of one to two years. It should also be borne in mind that, when introducing a new, more efficient technology, a significant part of the traffic needs to be processed with the older technology during a certain transition period and the advantages of the new technology will only be partly applied. Taking the advantage

of increased spectral efficiency due to new technologies, won't be enough to compensate for the rapid increase in mobile data traffic [26].
The following Section introduces three approaches how to tackle the challenge of increasing data traffic in mobile networks. The first two approaches are each a variable in Shannon's equation (number of antennas and transmission power) and the third approach presents alternative technologies which work complementary to mobile networks.

2.3.1 Increasing amount of cell towers

The huge growth in data traffic over the last year has meant that traditional macro-cell deployments no longer offer sufficient capacity to support the needs of their users [26]. In this subsection, we introduce two possibilities to meet traffic and data rate demands through increasing the amount of cell towers. On a high level, the key options to raise network capacity are densifying the macro cell network and complementing the macro cells with small cells, thereby creating a heterogeneous network.
The quality of mobile coverage of a geographic area is described by the coverage and capacity. Coverage means that a wireless connection is possible. The capacity provides information about the available data rate, how much data can be transmitted per second. The coverage is independent of the number of concurrent users, the capacity which is available in a radio cell, on the other hand, have to be shared with all active users simultaneously. The capacity of a network can be determined by compression, i.e. by reducing the radio cells with simultaneous reduction of the transmission power. A halving of the cell radius, for example, would quadruple the capacity approximately, with the required number of base stations also being quadrupled to cover a certain area[23] .
Each antenna can transmit only a limited amount of data simultaneously. The radio cell size depends on the expected number of users, the expected volume of data traffic and the topography. The greatest number of transmitters are therefore built in cities and towns, since cities and towns are more crowded and more people use mobile services on smaller space compared to rural areas. The cells which are supplied by the base stations have different sizes. Thus, the diameter of a radio cell in a rural area with several kilometers and in CBD may be less than 100 m [23]. The greater the demand for mobile voice and data services is the smaller the radio cells have to be designed and therefore the network of transmitters is more dense. In addition, smaller cells generally require a lower transmission power of the antennas and the mobile device, since the range is smalle [23].
Theoretically, reducing the size of a macrocell increases the cell's capacity. This sounds very promising for the tackling the problem of increased demand of data traffic. Reality however has shown that reducing the size of a macrocell is not that easy. For an efficient mobile network, the location of the base station is crucial. The facts that suitable locations are rare and that against every building application complaints have risen, makes it increasingly difficult and expensive for mobile providers to reduce the macrocells size. Further, in cities which already have a high density of macrocell base station, the densification of the network may be limited by high inter-cell interferences.
On the edge of two mobile cells interferences occur. To avoid or minimize the resulting interference, Inter-cell Interference Coordination (ICIC) is defined in 3GPP release 8 as an interference coordination technology used in LTE systems. It reduces inter-cell interferences by having User Equipments (UE), at the same cell edge but belonging to different cells, using different frequency resources. Base stations generate interference information for each frequency resource (RB), and exchange the information with neighbour base stations through X2 messages. Then, from the messages, the neighbour stations can learn the interference status of their neighbours, and allocate radio resources (frequency, Tx power, etc.) to their UEs in a way that would avoid inter-cell interference [7].

An alternative network structure of densifying the macro cell network is referred to as a Heterogeneous Network (HetNet)structure as the coverage area of the macro cells and local cells overlap [23]. By complementing the macro cell with micro cells, pico cells and femto cells, the per-user capacity and rate coverage will be increased.

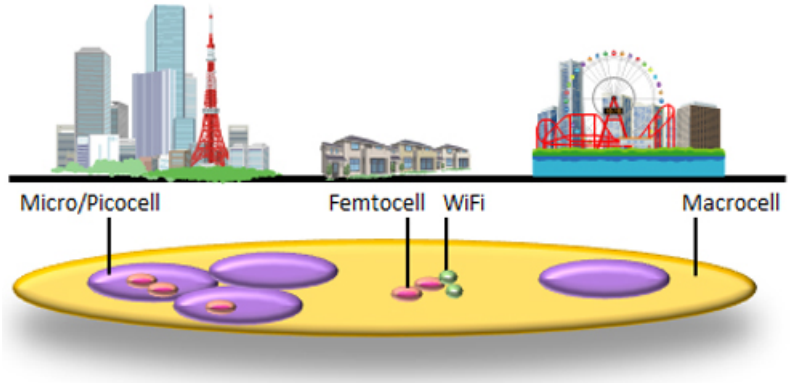


Figure 2.8: Heterogeneous Network utilizing mix of macro, pico, femto and WiFi

Consider the heterogeneous cellular system illustrated in Figure 2.8. This cellular system consists of regular deployment of macro base stations that typically transmit at high power level, overload with several pico base stations and femto base stations, which transmit at substantially lower power levels [4]. The small cells are deployed to improve capacity in crowded hot-spots areas and eliminate coverage holes in the macro-only system. As it has been mentioned, a deployment of a macro base station needs a careful network planning, the deployment of small cell base station can also be ad hoc. In addition, due to their smaller physical size and transmit power, it is easier to acquire sites for the small cell base stations. The main benefits of heterogeneous network solution are enriched capacity and guaranteed spot-free coverage. The overlay of small cells onto existing networks ensures high-speed, high-capacity communications for specific hot-spot area. According to [4], this is especially desirable in densely-populated areas and business districts that need to provide reliable communication services to support an increasingly diverse and heavy data range of applications.
ICIC was introduced to deal with interference issues at the cell-edge. While the basic ICIC scheme may be beneficial in a pure macro network environment, ICIC can not deal with interferences in a heterogeneous network environment where several small cells are located in the coverage area of a single macro cell [23]. 3GPP introduced in release 10 enhanced-ICIC to deal interference issues in heterogeneous networks. With this scheme, the macro cell coordinates with the small cells as to which of its subframes it leaves empty in the time domain. The small cells then use the empty macro-cell subframes for their own data transmission and thus avoid interference from the macro-cell in those subframes [4]. Again, a balance has to be found between the gain of reduced interference and the loss of transmission capacity in the macro cell and the small cells, since in the empty macro cell subframe no data transmission will occur.

2.3.2 Increasing radiation level

The second approach to engage the challenge of the increasing mobile data traffic would be to check and adjust the regulation of radiation levels emitted by mobile phones and/or the network infrastructure. This section focuses on the radiation of base stations of the network infrastructure. But before discussing current regulations in Switzerland and possible changes, this section will provide some basic definitions and explanatory details about radiation.

Radiation can be defined as "the transmission of energy from a body in the form of waves or particles"[12] or as "energy emitted in the form of waves (light) or particles (photons)"[18]. So all sources of light or energy are emitting radiation. The electromagnetic spectrum starts with radio waves, goes over to visible lights and ends with gamma rays. Another differentiation has to be done between ionizing radiation and non-ionizing radiation (NIR). The first type "has enough energy to strip electrons off of atoms"[12], the second doesn't have that much energy. NIR is emitted by radios, microwaves, mobile phones and base stations [13].
In the field of mobile communication radiation is emitted by mobile phones and base stations. The level of emitted radiation depends on the cell in which the mobile phone is used. In a macrocell (see section 2.2.2) the radiation level of the base station is low, but the mobile phone emits more radiation, in a microcell it is the other way around.
Other elements, which have an influence on the overall radiation of mobile communications, are the number of network providers, the numbers of users, their data traffic, as well as the topography of area in which the mobile communication takes place. The radiation level is regulated individually by each country, but is often based on the recommendations of the "International Commission on Non-ionzing Radiation Protection (ICNIRP)".
Table 2.2 shows the limits of the NIR regulation (german: NISV - Verordnung über den Schutz von nicht-ionisierender Strahlung of 12/1999) in Switzerland is 10x times higher than recommended by ICNIRP. About 6'000 (of the 15'000 base stations) already reached the limits defined in the NIR [25][26].

Table 2.2: Overview of radiation regulations [21]

	Basis of regulation	**NIR limits in V/M**		
		900 MHz	**2100 MHz**	**mixed**
Switzerland	ICNIRP and additional precautionary principle: Verordnung über den Schutz vor nicht-ionisierender Strahlung (NIS-V)	4	6	5
Germany	Bundesimmissionsschutzgesetz (BImSchG)	41	61	-
France	ICNIRP: Décret Nr. 2002-775, 3.5.2002	41	61	-
France (Paris)	Nouvelle charte parisienne de la téléphonie	5	5	5
Austria	ICNIRP: ÖVE/ÖNORM E 8850	41	61	-
Italy	DECRETO MINISTERIALE n. 381, 10.9.1998	6	6	6

The current NIR regulations in Switzerland are based on the definition of the site, the used frequency and methods for measuring exposure limits. One site can consist of antennas/base stations on different buildings, as long as they can be considered as one group of antennas. Different frequency ranges have different limits. Antennas, which are using 900 MHz, are allowed to emit 4.0 V/m, with 1800 MHz the allowed radiation level is 6 V/m and all other antennas are only allowed to reach 5 V/m. But if more then one antenna is installed on one building or several buildings, which are considered as one site, those limits apply for all antennas together and not only one. So if the different antennas belong to different network providers they have to collaborate and have to find an agreement which ensures they stay together in the allowed ranges [25]. Additionally, the measurement of the radiation differs from country to country. For example if the radiation is measured close to a base station or in the direct line of the main beam the radiation level is higher, then if the measuring is done from a greater distance [22].
There are different adjustments possible which would result in an increased radiation level, but would also give more flexibility to network providers at low cost. The first adjustment could be to change the definition of a site. So a site could consist only of antennas on the same building, it then won't matter any more how close the antennas are to each other. This would effect especially urban areas. Another positive effect in terms of flexibility

would be if only antennas from one network provider are defined as one site. In the worst case scenario the radiation level would then increase by the factor 2 or 3, assuming that all three network providers of Switzerland (Swisscom, Salt and Sunrise) would reach the radiation level limits. The third possible adjustment would exclude all antennas which use less then 6 watt as an effective radiated power (ERP). This would increase the possibility of installing antennas to realize micro and picto cells. That would decrease the radiation of base stations, because the distance between the base station and the mobile phone user is decreasing [26].
Another possibility would be that there is only one radiation level limit in the future, independent of the used frequency. The effects of different thresholds are listed in an study of the Swiss Federal Council [26]. It would be also possible to introduce the possibility to get an special permit to exceed the defined limits in some cases, e.g. in the CBDs of cities [26].

2.3.3 Integration of other mobile technologies

One option to bypass the mobile signal is WiFi Calling. WiFi Calling is a service which allows users to make voice calls without cellular service. This service is preinstalled in many new smart phones. In WiFi Calling a Mobile Node is connecting to a Access Point (AP) with the wireless interface in the frequency range of 2.4 GHz and 5GHz. Figure 2.9 shows the AP connecting with the IP Network to the Uma Network Controller (UNC) [11]. The UNC is connected to the MSC which routes the calls trough the cellular network.

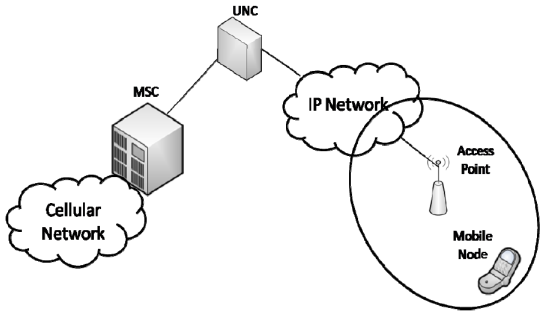


Figure 2.9: Wifi-Calling [11]

Another possibility is Whatsapp Calling. In 2015 Whatsapp activated a new feature for calling [28]. Every of the one billion users is now able to call through Whatsapp. If the smart phone is connected to an wireless network the call can be made without cellular service.
There are dozens of other applications like Whatsapp, for example Skype or face time. Many of those applications are build on Voice over IP (VoIP). In VoIP Session Initiation Protocol (SIP) is used for signaling and Real-Time Transport Protocol (RTP) for transfering the voice data. Swisscom wants to switch off the Integrated Services Digital Network (IDSN). After that all voice communication is done by VoIP.
To unload the cellular network the coverage of WiFi networks have to be increased. Many shopping centers, railway stations and universities have a area-wide coverage. With the integration of other mobile technologies, the frequency and cell size changed. The integration brings the possibility to integrate an existing installation, which is very cheap. But this integration brings a change of requirements for WiFi deployments. A few years ago the area-wide coverage was the aim. Because of new possibilities and subscribers in the last years, the importance of bandwidth and stability has grown. There is one big obstacle, the handover between the Access Points is no problem, but with a handover between

the WiFi and the cellular network the call fails. Therefore, this is necessary being connected to the same WiFi network during the whole call. In a university, shopping center or railway station it is possible, but what if the call takes longer. Public WiFi's would be a solution. A Public WiFi covers a wide range of a CBD. For example in Lucerne the public WiFi covers a surrounding of 300m around the railway station.

2.4 Effects of Radiation on Humans

After defining basic terms about radiation in section 2.3.2 this section will provide more details on effects of radiation on the human body. As previously stated it is always important to consider the various sources of radiation, which are emitted to the human body like radio, television or micro ovens besides the mobile phone and base stations.
In terms of effects of radiation it has to be differentiated between thermal or short-term effects and non-thermal effects, which are also considered to have to long term effects on the human body. Thermal effects are caused by energy transported by the radiation of mobile phones. This energy leads to a temperature rise in the brain or any other organs of the body due to the fact that the human body consists of 70% of water, which is also known as the microvave effect [15][14]. According to an article cited in [14] possible symptoms originated from the thermal effects could be the occurrence of: "headaches, sleep disorders, poor memory, mental excitation, confusion, anxiety, depression, appetite disturbance and listlessness"[14].
Non-Thermal effects aren't perceived immediately, but they sum up over time and could be even more severe after a few years [15]. Cell phone and base station radiation are associated with a higher risk of cancer and brain tumors. Furthermore in [3] it is claimed that it also weakens the blood barrier, reduces sperm counts and increase the probability of a miscarriage. One important fact is also that the conducted long term studies so far didn't consider modern usage patterns, since today the usage time of the mobile phone is much higher than a few years ago.
The World Health Organisation (WHO) is classifying radio-frequency electromagnetic fields as possibly carcinogenic to humans (Group 2B), a category used when a causal association is considered credible (along with e.g. coffee) [29]. Several other studies were conducted to check the results of the studies mentioned in the previous section. The Mobile Telecommunications and Health Research Programme states that "neither of the studies identified any association between exposure and an increased risk of developing cancer"[17]. Another paper reported "methodological limitations in available studies, primarily recall bias, reversed causality, confounding, and selection bias, prevents conclusions about causal effects of RF exposure on the studied health outcomes in children and adolescents"[9] and is also criticizing the number of available studies to get to an profound conclusion about the effects of cell radiation [9].
Another study tested differences of perception, based on humans character traits, for example if they would see themselves as sensitive or not. The authors of the study used "open provocation and double-blind tests to determine if sensitive and control individuals experience more negative health effects when exposed to base station-like signals"[8]. The result was that they didn't find any difference between the sensitive and control group people in the double-blind test or in physiological measures. But the people which characterized themselves as sensitive had the subjective feeling that the antennas have an negative impact on their well-being.
Unfortunately, long term studies aren't available which are also taking into account that the use of mobile phones has increased over the last few years. One attempt to overcome this short coming is a cohort study, which started in 2010. The goal is the "study of mobile telephone users (ongoing recruitment of 250,000 men and women aged 18+ years

in five European countries - Denmark, Finland, Sweden, The Netherlands, UK) who will be followed up for 25+ years"[24].

2.5 Cost and Benefits Analysis

To compare the three approaches presented in section 2.3, this section introduces an extensive cost calculation model to analyze the cost driver and their impact on the mobile network. Further, the appreaches are compared to each other regarding five different perspectives.

2.5.1 Cost calculation model

To calculate the costs of a mobile network and analyze the impact of changes on the cost drivers (e.g NIR regulation), a simplified but still realistic model is used [21]. Figure 2.10 shows the cost calculation model which can be divided in three steps. The first step is to identify the dimensions of the network. By determining the dimensions of a network, mobile providers are able to calculate the number of antenna sites required to satisfy the customer demands.
The first input factor is the **Regulatory Framework**. This input concerns with any restrictions and regulations which are in place. Differences in NIR regulation have a significant impact on mobile network costs. Strict regulations lead to the fact that fewer carriers per site can be built or lower coverage can be attained. To provide the coverage and capacity demanded by the customers, mobile network providers need to build more antenna sites. Consequently, NIR exposure limits have the higherst impact on mobile network costs compared to other cost drivers [21].
The second input factor is the **Market Information**. This input sets the dimension regarding the country coverage or the consumer demand for mobile data (capacity) and the smartphone penetration. The third input factor is **Technical Information**. This input includes all technical parameters and characteristics of the mobile network for example, used frequency bands, interference margin or sensitivities, distribution of macro- and micro-cells and the distribution of the antenna sites.
The last input factor for dimensioning the network **Country-specific Characteristic** parameters. This includes all detail and basic data of country such as the size of the population and information regarding topography.
Further, the distribution of urban, suburban and rural areas is taken into account. Since in most rural areas the mobile network only satisfies the basic coverage but no capacity demand, the costs to meet increased demand in rural areas are expected to be higher than in urban, respectively suburban regions. Also, border regions and mountains are considered as an input, as well as the total length of tunnels and railways. Border regions are identified as one of the most significat cost drivers [21] since mobile providers have to reduce interaction with the mobile providers in neighbouring countries, such as a restricted available frequency spectrum. Therefore, antennas in border regions have a stronger tilt and smaller range of coverage. To fulfill the demand of mobile services in border areas, more antennas are required. In case mobile providers of the neighbouring country have higher NIR limits, mobile phones would connect to the neighbouring mobile network. To prevent this situation and keep the customers in the local mobile network, the mobile providers have to operate more antennas. Mountains are another cost driver for mobile networks. The size of the region and the characteristics of the mountains determine the number of antenna sites required to cover the area. The shape of mountains leads to many blind spots which have to be covered by additional antennas. However, not only the characteristics of the mountains lead to the cost driver, addional costs are mainly

driven by the costs of connecting antenna sites. [21] Tunnels are further cost drivers for mobile networks. Tunnels exceeding a given length cannot be covered with external antenna sites and require additional antennas inside the tunnels. As a result, long tunnels in a country have an impact on the mobile network costs. Also the coverage of railway lines influence the mobile network costs. Capacity demand along railway lines increases in temporary burst when a train passes carrying a high number of simultaneous users of mobile devices. This high capacity demand can be satisfied through more antenna sites along the railway tracks.

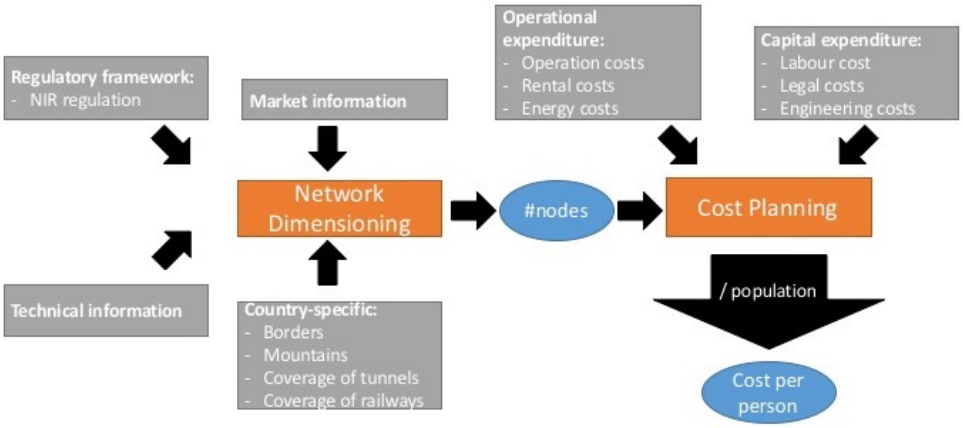


Figure 2.10: Cost Calculation Model based on [21]

As a second step in the modeling is to calculate the cost planning. Cost planning is based on two factors, the operational expenditures and capital expenditures. **Operation expenditure** includes all costs for operating the antenna sites. One component is rental costs of a location for a site. Another component is the energy costs for the operation the antenna. **Capital expenditure** is the second factor, it includes all costs for labour, legal and engineering. Constructing, operating and maintaining mobile networks give rise to labour and service costs.
As a third step, the cost per person could be calculated by dividing the cost planning through the number of customers. However, this measurement is not relevant to compare the presented approaches.

2.5.2 Costs and benefits comparison

The three approaches presented in section 2.3 are compared against each other. Table 2.3 shows the cost- benefit analysis amoung the three approaches. The analysis is split into 5 subcategories. The classes (very bad, bad, good and very good) define the impact of the approach regarding the given category and the future mobile traffic and data rate demands. Table 2.3 also includes the subcategory Non-Thermal. The authors decided to add it for completness, even no approach has a proven impact on the non-thermal effects. Since there is no study which proves that cell radiation has a non-thermal effect on the human body.
Increasing the amount of cell towers has a negative impact on costs. On the one hand, operational expenditure are increasing due to higher location rental costs, as well as higher energy costs. On the other hand, capital expenditure raises too. Constructing, operating and maintaining antenna sites give rise to labour and service costs. As it has been mentioned, borders are a big cost driver for mobile network providers. With a higher number of antenna sites the demand for mobile traffic and data can be satisfied in border areas and therefore, it has a positive impact. To provide coverage in mountain areas, mobile network providers have to distribute their sites around the mountain. Several

characteristics of the mountains, in addition to height, have a significant influence on the distribution. An increased amount of cell towers is therefore, inevitable to provide good coverage in mountain areas. To provide coverage in tunnels longer than 1.4 km additional antennas inside the tunnel are required [21]. Consequently, an increased amount of cell towers has a positive impact on coverage of tunnels. To meet the higher capacity demand along railway lines, an increased number of mobile antennas along railway tracks is required. As a result, more cell towers have a good impact on coverage of railways. As mention in Section 2.3.1, there is inevitably some level of interference from the signal from the other cells which use the same frequency. In a more densified macro-cell network interferences are even stronger. In heterogeneous networks eICIC reduces the interferences by having the macro mute some of its subframes. With a higher number of antennas, the macro cell has to sacrific more resources and hence it has a negative impact. More mobile antennas automatically increase the radiation level.

Table 2.3: Cost- / Benefit Analysis

	Cell Towers	Radiation Level	Technology Integration
Costs			
Operational expenditure	- -	-	+ +
Capital expenditure	- -	+ +	+ +
Regulatory			
NIR regulation	0	+ +	0
Country-specific			
Borders	+	+	0
Mountains	+	0	0
Coverage of tunnels	+ +	- -	0
Coverage of railways	+	0	+ +
Technical			
Interference	-	-	+ +
Health			
Thermal	-	-	0
Non-Thermal	0	0	0
Total	+	+	+ +

- -: very bad, -: bad, 0: no impact, +: good, + +: very good

Increasing the radiation level has an overall positive impact on the costs. As higher transmission power has a higher consumption of electricity and hence a negative impact on energy costs. However, loosen the NIR regulations has a very positive impact on capital expenditures. In general the bureaucracy effort for the mobile providers decrease, e.g. gaining a authorization for operating a site and transmitting will be less challenging due to less strict regulations. Further, authorizations are subject to litigation, which causes legal cost and supervision costs. Increasing radiation level will therefore, also result in less legal costs. Having a higher radiation level is also helpful at borders to counter the signals from a foreign operator which already sends with a higher transmission power. Higher radiation level does not have any impact on mountains, as the main challenge constructing, operating and maintaining a mobile network in a mountain area is to provide full coverage. Increasing the NIR limits does not solve the problem of blind spots. Regarding interferences, higher transmission power leads to the same problem than increased number of antennas, therefore it is graded as negative.
Integration of other technologies have a positive impact on the costs. Many of the alternative antennas, e.g. WiFi at University, is not constructed, operated and maintained by mobile network providers and hence there are no costs for the mobile network providers

although their network gets unloaded. As antennas for other mobile technologies all work as low power nodes, NIR regulation has no big impact on this approach. As the area of application for these alternatives are very specific, they also have no impact on borders, mountains and coverage of tunnels. Other mobile technolgies use odd frequencies and therefore no interferences occur when deploying them.
The total in Table 2.3 shows that integration of other mobile technologies is the only approach which has an overall total of very good. Integration of other mobile technologies is indeed very beneficial, however this approach can not solve the problem of increasing mobile data demand by it self. Integration of other mobile technologies is more complementing the other two approaches. Looking at the overall total of increaseing the amount of cell towers and increasing the radiation level, both approaches have a total of good. Both approaches are very beneficial with the same relative costs. If there has to be choosen one approach, increasing the radiation level would be prefered. It appears that increasing the radiation level is a more sustainable solution. Increasing the amount of cell towers requires cost intensive infrastructure, whereas increasing the radiation level can be initiated by loosen the regulation. Further, increasing the radiation level is also easier to revoke. The big disadvantage of increasing the radiation level is that it does not provide coverage in tunnels and mountain areas, therefore in these areas have to be built more antennas for a better coverage and capacity.

2.6 Summary

After presenting challenges of the growth of mobile traffic three approaches for solving those challenges were described in detail. The first approach is to increase the number of cell towers / base stations to satisfy the need for more bandwidth and a faster connection to watch e.g. more videos with the mobile phone. Advantages are that new base stations could be then specifically designed and deployed for the previous introduced technologies like LTE or the upcoming technology 5G. Disadvantages are that is very costly and the search for suitable locations takes a lot of time. Furthermore the finding and deployment process often involves law suits against those new base stations.
The second approach is about changing the regulations of the non-ionizing radiation (NIR), which includes the radiation which is emitted by cell phones and especially base stations. It has stated that the regulation in Switzerland are more strict than in other European countries as well as the recommended radiation level restrictions by the ICNIRP. In addition to that also the definition of the site and the measurement of radiation could be changed, so that the current base station could be used at an higher degree. At the moment 6'000 out 15'000 base stations already reached their radiation limit based on current regulations.
The integration of different technologies was described as an third approach to tackle the faced challenges. This would include an decrease of the data traffic, which is transferred over mobile networks, and an increase of the data traffic over various WiFi networks, e.g. in shopping malls or office buildings. Advantages are that all technologies are already available and in use and the radiation level wouldn't be increased. Examples for disadvantages are that the handover between those technologies e.g. in an voice call has to be improved, like it has been done for doing voice calls in trains over the mobile network only.
For all approaches effects of the non-ionizing radiation on humans should be taken into account. In general effects of radiation can be divied into thermal (short-term) and non-thermal (long-term) effects. Various singular studies have shown a broad spectrum of possible effects or symptons. But it is very difficult to generalize those results as other studies stated, especially for the influence of base stations isolated. Additionally due

to the increased usage of mobile phones there aren't any long term studies available to support or deny assumptions about long term effects.
In conclusion all the research questions are answered and it seems most suitable to use a combination of the different approaches to face the challenges and to satisfy today's and future needs of mobile phone users, at least according to the opinion of the authors. This would also take into account the specific situation of Switzerland with the border regions, mountains and tunnels.

Bibliography

[1] BAKOM, "Standorte von Sendeanlagen", https://map.geo.admin.ch/?topic=funksender, 2016, [Online, accessed 2016-10-20].

[2] M. R. Bhalla, A. V. Bhalla, "Generations of mobile wireless technology: A survey," *International Journal of Computer Applications*, Vol. 5, No. 4, pp. 26–32, August 2010.

[3] N. Cherry, "Health effects associated with mobile base stations in communities: the need for health studies," http://bit.ly/2hoABR2, [Online, accessed 2016-11-15].

[4] D. X. Chu, D. D. Lopez-Perez, P. Y. Yang, D. F. Gunnarsson, *Heterogeneous Cellular Networks: Theory, Simulation and Deployment*. New York, NY, USA: Cambridge University Press, 2013.

[5] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, [Online, accessed 2016-10-20].

[6] H. Claussen, L. T. Ho, L. G. Samuel, "An overview of the femtocell concept," *Bell Labs Technical Journal*, Vol. 13, No. 1, pp. 221–245, March 2008.

[7] M. M. Do, H. J. Son, "Interference coordination in lte/lte-a (1): Icic," http://www.netmanias.com/en/post/blog/6391/icic-interference-coordination-lte-lte-a/interference-coordination-in-lte-lte-a-1-inter-cell-interference-coordination-icic, [Online, accessed 2016-10-20].

[8] S. Eltiti, D. Wallace, A. Ridgewell, K. Zougkou, R. Russo, F. Sepulveda, D. Mirshekar-Syahkal, P. Rasor, R. Deeble, E. Fox, "Does short-term exposure to mobile phone base station signals increase symptoms in individuals who report sensitivity to electromagnetic fields? a double-blind randomized provocation study," *Environmental health perspectives*, Vol. 115, No. 11, pp. 1603–1608, July 2007.

[9] M. Feychting, "Mobile phones, radiofrequency fields, and health effects in children – epidemiological studies," *Progress in Biophysics and Molecular Biology*, Vol. 107, No. 3, pp. 343 – 348, September 2011.

[10] Gartner, "Gartner says worldwide smartphone sales grew 3.9 percent in first quarter of 2016," http://www.gartner.com/newsroom/id/3323017, [Online, accessed 2016-10-18].

[11] S. F. Hasan, N. H. Siddique, S. Chakraborty, "Femtocell versus wifi - a survey and comparison of architecture and performance," in *Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on*, May 2009, pp. 916–920.

[12] D. Howard, "What is radiation? - definition, causes and effects," http://study.com/academy/lesson/what-is-radiation-definition-causes-effects.html, [Online, accessed 2016-10-28].

[13] International Commission on Non-Ionizing Radiation Protection, "Non-ionizing radiation," http://www.icnirp.org/en/home/home-read-more.html, [Online, accessed 2016-11-12].

[14] M. Kaushal, T. Singh, A. Kumar, "Effects of mobile tower radiations & case studies from different countries pertaining the issue," *International Journal of Applied Engineering Research*, Vol. 7, No. 11, pp. 1252–1255, May 2012.

[15] N. Kumar, G. Kumar, "Biological effects of cell tower radiation on human body," *ISMOT, Delhi, India*, pp. 678–679, 2009.

[16] J. Marcus, *Study on impact of traffic off-loading and related technological trends on the demand for wireless broadband spectrum : final report.* Luxembourg: Publications Office, 2013.

[17] MTHR Programme Management Comittee, "Mobile communications and health research programme report 2012," http://www.mthr.org.uk/documents/MTHRreport2012.pdf, [Online, accessed 2016-11-03].

[18] National Aeronautics and Space Administration, "What is radiation? - definition, causes and effects," http://imagine.gsfc.nasa.gov/science/toolbox/emspectrum1.html, [Online, accessed 2016-11-12].

[19] NEC Corporation, "Nec on heterogeneous networks," http://se.nec.com/enSE/global/solutions/nsp/ltesc/hetnet.html, [Online, accessed 2016-10-16].

[20] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the metis project," *IEEE Communications Magazine*, Vol. 52, No. 5, pp. 26–35, May 2014.

[21] PricewaterhouseCoopers AG, "Mobile network cost study," https://asut.ch/asut/media/id/94/type/document/st_pwc_mobile_network_cost_20130904.pdf, 2013, [Online, accessed 2016-10-19].

[22] S. H. Saeid, "Study of the cell towers radiation levels in residential areas," in *International Conference on Electronics and Communication Systems 2013*, Rhodes Island, Greece, July 2013, p. 87.

[23] M. Sauter, *From GSM to LTE-Advanced - An Introduction to Mobile Networks and Mobile Broadband*, 1st ed. New York: John Wiley & Sons, 2014.

[24] J. Schüz, P. Elliott, A. Auvinen, H. Kromhout, A. H. Poulsen, C. Johansen, J. H. Olsen, L. Hillert, M. Feychting, K. Fremling, M. Toledano, S. Heinävaara, P. Slottje, R. Vermeulen, A. Ahlbom, "An international prospective cohort study of mobile phone users and health (cosmos): Design considerations and enrolment," *Cancer Epidemiology*, Vol. 35, No. 1, pp. 37 – 43, February 2011.

[25] Schweizer Bundesrat, "Verordnung über den Schutz vor nichtionisierender Strahlung (NISV)," https://www.admin.ch/opc/de/classified-compilation/19996141/index.html, 1999, [Online, accessed 2016-11-12].

[26] ——, "Zukunftstaugliche mobilfunknetze: Situationsanalyse," https://www.bakom.admin.ch/dam/bakom/de/dokumente/zukunftstauglichemobilfunknetze.pdf, 2015, [Online, accessed 2016-10-19].

[27] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, Vol. 5, No. 1, pp. 3–55, January 2001.

[28] J. M. Vanerio, P. Casas, "Whatsapp calling: a revised analysis on whatsapp's architecture and calling service," in *Proceedings of the 2016 workshop on Fostering Latin-American Research in Data Communication Networks*. ACM, 2016, pp. 13–15.

[29] World Health Organization, "Electromagnetic fields and public health: mobile phones," http://www.who.int/mediacentre/factsheets/fs193/en/, [Online, accessed 2014-10-29].

Chapter 3

Comparing Blockchains

Patrick Dueggelin, Camilla Gretsch, Daniel Oertle

With the success of Bitcoin numerous blockchain technologies have emerged and gained attention of developers as well as investors. With this many new technologies we try to introduce and compare six different emerging projects that utilize elements of blockchains. First we give a short introduction into the blockchains with Bitcoin and Ethereum as examples, then we introduce Monero, Rootstock, IOTA, BigchainDB, Arcade City and La'Zooz. We show an overview and compare these eight technologies in their project scope, performance and production readiness. We came to the conclusion that many new technologies still have to proove themselves in the real world.

Contents

3.1 Introduction

The blockchain is simply a data structure which consists of several blocks. The blocks are chronologically linked together while each contains the hash of the previous block. The white paper "Bitcoin: A Peer-to-Peer Electronic Cash System" shows the concept of using blockchains as a distributed database. This is why the term blockchain is often mentioned in direct relation to Bitcoin and cryptocurrencies. Bitcoin allows payments over the internet without a trusted third party and is the first cryptocurrency based on the blockchain technology, it was released eight years ago in 2008. The technology behind blockchains used as a cryptocurrency is still young and improves continually. Today over 600 different cryptocurrencies exists and the list is still growing.
The blockchain technology is also usable in other areas than financial institutions. In particular everywhere where data has to be stored decentralised in a secure way. In a notaryship to verify authenticity of documents. In the music industry where the blockchain could contain music rights ownership informations. Or just to store any kind of data online like Storj does it. Storj is blockchain based peer-to-peer cloud storage platform. It can be said that the blockchain technology can be used in many different ways.

3.2 Blockchain technology

This section gives a brief overview of the blockchain technology. At first is the historical background shown, followed by an explanation of the core concept of blockchains. For this purpose all main terms used in connection to this topic are defined in here. The concept of the blockchain technology is explained through how Bitcoin and Ethereum work.

3.2.1 Bitcoin

The Bitcoin design paper was published in 2008 by Satoshi Nakamoto. The main idea of the paper is an electronic payment system based on cryptographic proof [3], thus payments over the internet can be made directly from one party to an other one without using a trusted third party service (financial institution) [3, 4]. With this paper the idea of the cryptocurrency was born. Bitcoin was the first ever published cryptocurrency and is still the biggest one [1].

3.2.1.1 Description of the System

Network Bitcoin is a peer-to-peer system, in which every party is represented by a node [3]. When one node sends an amount of Bitcoin to another node, it is called a transaction [5]. Bitcoin is not only the name of the cryptocurrency it is also the name of the currency itself. A new transaction is always broadcasted to all nodes in the network. The validation of the transaction is checked and only then it is copied into a block. A new block has to be mined before it is added to the blockchain, this is done by proof-of-work and only after this step the transaction is executed [3, 5]. This is just a brief overview of the Bitcoin network, which is explained more detailed in the further sections.

Transaction Every node has two unique keys, one is private and one public. The private key can only be seen by the owner and is used to create a digital signature. The public key is seeable for every node in the system and is used to verify the signature. It can also be seen as the address of the node [5, 6]. A valid transaction consists of the transactions content, the address (public key) of the receiver, the signature and the public key of the sender, and a reference to previous transactions [5].

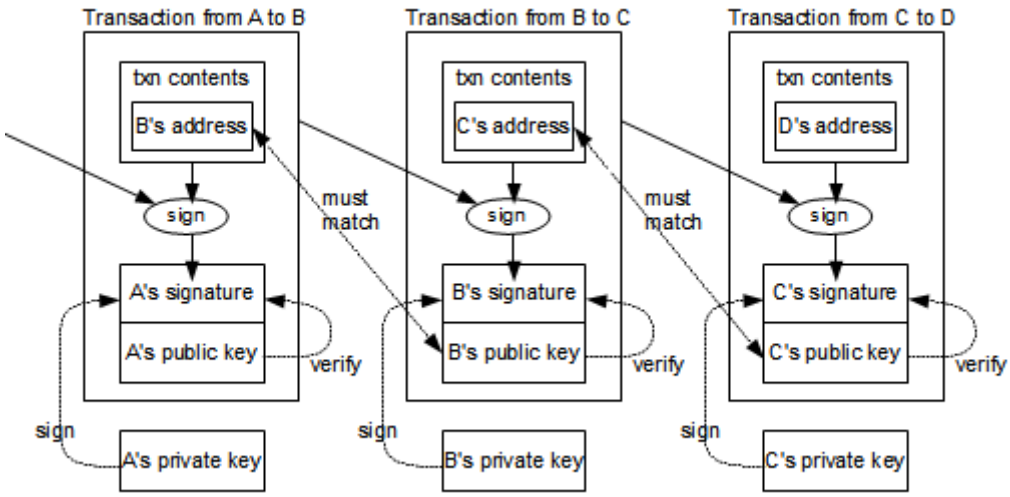


Figure 3.1: three valid transactions [5]

In Fig. 3.1 node B transfer coins to node C. The digital signature is created by the private key of B. The other nodes in the system can verify the signature with the public key of B. This ensures that B is truly the owner of the Bitcoins [4, 5]. The transaction also contains a hash of the transaction from A to B, which signifies that B received Bitcoins from A and is now spending them to C. All this transactions, linked with previous transactions, can be considered as a transaction chain [5]. With the transaction chain the other nodes are able to see if B is really in possession of these Bitcoins and where they come from. It also ensures that B can not spend Bitcoins he is not in possession of or double-spend them [4]. A transaction can also have more than one previous transactions [6]. As shown in 3.2, each "input" are transactions with received Bitcoins including the address of the sender. The "output" are transactions with the spent Bitcoins and the address of the receiver [5]. As soon as the Bitcoins are transferred, the "inputs" are locked therefore that the coins can not be double-spent [5].
With the transaction chain the nodes can keep track of the balances of every single node in the system by adding up all the unspent "inputs " minus the "outputs "Even for the own balance the node has to iterate through all transactions referenced to his public key [6].

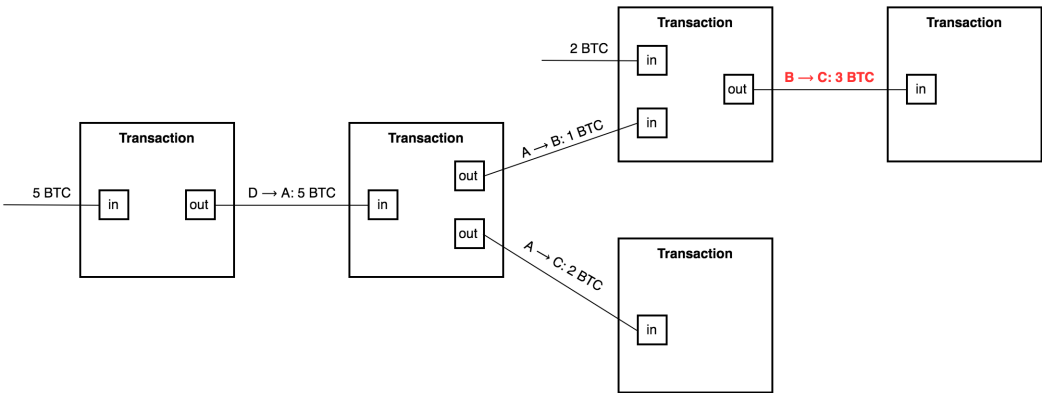


Figure 3.2: transaction chain

Blockchain Shown in Fig. 7.1.2.1 the blockchain consists of several chronologically linked blocks which contain transactions, a nonce value and a hash to the previous block [3, 4]. So in the blockchain every transaction ever made is recorded.

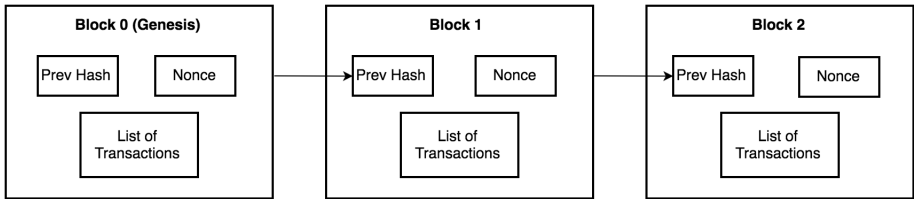


Figure 3.3: Blockchain, where Genesis is the very first block

Mining ad Proof-of-Work So far explanation is given to which criteria a valid transaction must be fulfilled. Mining is the process of verifying new transactions and adding them to the blockchain [3]. New transactions are broadcasted to the network, where they first land in a pool of unconfirmed and unordered transactions. Since the nodes can not be sure whether the order in which the transactions were received is the same as they were generated [4]. Special nodes, called miners, gather unconfirmed and unordered transactions and copy them into a block [4]. Transactions packed into one block are considered as to have happened at the same time. The miners broadcast their block, containing new transactions in an ordered way, as a suggestion for the next block in the chain. Though problematic is when several miners are suggesting their block simultaneously but it is solved through proof-of-work [4].
Each miner has to solve a cryptographic hash (SHA-256) by random guessing, which needs a lot of computing power [4]. The solution is written on the nonce value and then the block broadcasted to the network. The other miners verify the solution of the hash function and when the majority accepts it, it is added to the chain. The majority of the network is determined by the nodes with the most computing power [3, 5].
It is still possible that several blocks are broadcasted to the network at the same time, if so, different branches are built. The nodes are working on the first block they receive, therefore different nodes are working at the end of several branches in the chain [3]. In general all nodes switch to the longest chain so the blockchain is stabilized again. The longest chain is considered the honest chain because it has the greatest proof-of-work effort. If the majority of computing power is controlled by honest nodes, then the honest chain is the fastest-growing one [3]. In fig. 3.4 the blocks 3 and 5 are part of the shorter chain, therefore they are going to be abandoned and the transactions in these blocks are going back to the pool of unconfirmed and unordered transactions where they have to wait till they are added to a new block [3].

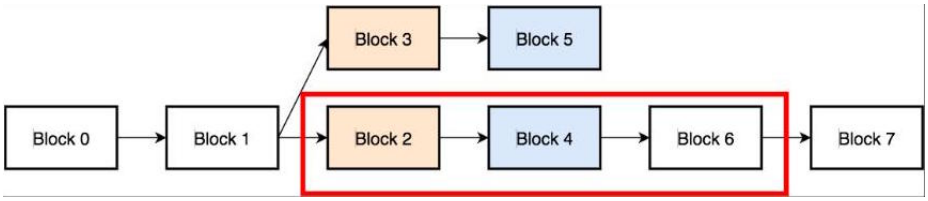


Figure 3.4: longest (honest) chain is controlled by the majority of computing power

Mining reward Bitcoins are rewarded to the miners for their efforts. Right now the mining reward is at 12.5 Bitcoins but it is halved every 210000 blocks [7]. Every block has an extra transaction which contains new coins, credited to the owner of the block. In this way new Bitcoins are distributed to the network. Additionally the miners receive transaction fees, which are payed by the nodes with a sending transaction in this block [3].

3.2.1.2 Performance

In average a new block is mined every ten minutes [3]. Due to the Moore‘s law the computing power will increase over the years [10]. To compensate this, the cryptographic hash difficulty is increasing too, so that the block time is constantly about ten minutes [3].
The general rule is, that a node has two wait six new blocks till he can be sure that his transaction is in the blockchain. Thus the blocktime is ten minutes and it takes six block, it can be said, that the confirmation time for a transaction can take up to one hour (= 10 min * 6 blocks) [6].
The theoretical maximum of handling transactions per seconds (tps) is at seven. But this could only be possible when all transactions in a block has minimal size. When the transactions only have one input and output. In reality the transaction size is larger and the transactions per seconds are about three to four [8].

3.2.1.3 Attacks and Challenges

Several different possible attacks for the Bitcoin systems exist. One of them, the 51% attack, is explained here. When one node reaches 51% of the computing power, it can build its own chain faster than the network [9]. Since its chain is longer than the honest chain, all nodes in the network consider now that chain as the new honest chain and switch to it. During this time the attacker can modify the transaction order. He can reverse transactions he already sent, what can lead to double spending or prevent transactions from getting confirmed [9]. Therefore the mining reward is supposed to encourage nodes in staying honest. An attacker can choose between using his own computing power to attack the network or to generate new coins by mining blocks [3].

3.2.2 Ethereum

In 2013 Vitalik Buterin published the Ethereum white paper, with the propose: “a blockchain with a built-in Turing-complete programming language, allowing anyone to write smart contract and decentralized applications where they can create their own arbitrary rules for ownership, transaction formats and state transaction functions“ [11, p.13]. Ethereum is based on the same technology as Bitcoin but is modified in some points and owns additional features, for example smart contracts [12].

3.2.2.1 Description of the System

This section explains how Ethereum works. The techniques are the same as in Bitcoin but aren‘t explained precisely here. For further details regard section 3.2 Blockchain technology.

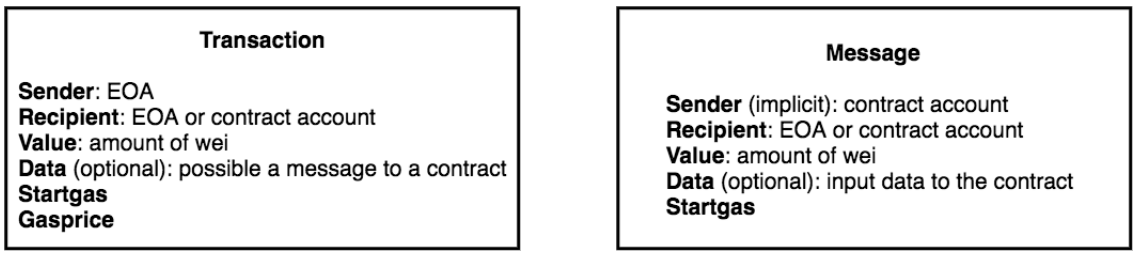
Network Ethereum is a peer-to-peer network in within every client is represented by a node [12]. The basic unit of Ethereum are accounts. With the state transition function Ethereum keeps track of every account and additionally, values and information can be transferred between them [11]. Miners verify new transactions and group them together into a block. Blocks have to be mined by proof-of-work before they are broadcasted to the network and the transaction must be executed [12]. Public and private blockchains exist in Ethereum. Public blockchains work in the same way as in Bitcoin. Every node in the Ethereum network can read and send transaction. The private Blockchains normally only have write and read permissions to one organization. Furthermore every node can write smart contracts and runs the Ethereum Virtual Machine [12].

Smart Contract and Ethereum Virtual Machine (EVM) Smart contracts are basically physical contracts written in a computer language. The purpose of smart contracts is to formalize and ensure relationships over computer networks [13]. Every user in Ethereum can create his own smart contracts. The EVM is used to built and run these smart contracts. It can execute codes of any complexity, this means Ethereum is turing complete [12].
Users can create new contracts like the SimpleStorage contract in fig. 3.5. To do so the user writes the SimpleStorage contract in Solidity, which is an Ethereum high-level language. Then the contract is compiled by the EVM-compiler to bytecode and uploaded by the user to the blockchain. The EVM executes the bytecode. Every node in the network runs the EVM and executes the instructions [12].

```
pragma solidity ^0.4.0;

contract SimpleStorage {
    uint storedData;

    function set(uint x) {
        storedData = x;
    }

    function get() constant returns (uint retVal) {
        return storedData;
    }
}
```

Figure 3.5: Smart contract written in Solidity; used to modify or retrieve values [14]

Accounts Two different types of accounts exist. One is the externally owned account (EOA), which is a personal account owned by an user. It contains an ether balance and is controlled by the user‘s private key [12]. Ether is the name of the currency of Ethereum. The other one is the contract account, which is controlled by the smart contract code and also contains an ether balance [12]. In Ethereum, transactions are signed data packages which can only be sent by EOAs but received by both account types. The fig. 3.6a shows what a valid transaction must contain. With a transaction, EOA can transfer coins to another EOA or trigger the code of contract accounts [11]. Messages are similar to transactions, as seen in fig. 3.6b, but are only sent by contract accounts. They are produced when the contract account code executes a “call“ function [12].



(a) Transaction **(b)** Message

Figure 3.6: Content of a transaction and message in Ethereum

When the code of a contract is triggered by a transaction or message then every instruction is executed by every node in the network. This is not very efficient since the code execution is redundantly parallel, but ensures the consensus of the network [12]. To protect the network from infinite loop attacks, senders have to pay a fee per transaction. The fee is determined by the values startgas and gasprice. Startgas is the limit of computational

step of code execution and is defined by the sender [11]. The fee per computational step is called gasprice. When the fee for a transaction is greater than (startgas * gasprice), it is called "runs out of gas". In this case all state changes are reverted, except the payment of the fees [11].

Mining The block validation process of Ethereum is similar to the one in Bitcoin. Blocks also contain the hash of the previous block, a nonce value and the list of transactions. Furthermore the blocks have the most recent state of accounts, block number and the hash difficulty included [12]. Ethereum uses proof-of-work to mine new blocks, with a hash function called Ethash. The miners are getting rewarded with constantly five ether and the fees for the transactions or messages depends on the gas price [12].

3.2.2.2 Performance

In Ethereum the blocktime is at about 15 seconds and it takes 12 blocks to be sure that the transaction is added to the blockchain. This leads to a transaction confirmation time of three minutes (= 15 sec * 12 blocks). The Ethereum network can handle 20 - 30 transaction per seconds. They even go one step further, they announced to increase the transactions per seconds to 10000 in the next release by using proof-of-stake instead of proof-of-work [12, 15].

3.3 Blockchain services

In this section five different services that are based on blockchain technology are shortly introduced. In the section "Description of the system" their main, technical and non-technical components are summarized. Then the performance data, such as the transactions per second that the network can handle, is discussed in the section "Performance". At last possibilities for attacks and other challenges are presented.

3.3.1 Monero

Monero is a privacy-focused cryptocurrency based on the CryptoNote protocol. It was launched on April 18, 2014 and recently (end of August 2016) gained the attention of the cryptocurrency community [35]. Currently, Monero ranks 5th in terms of cryptocurrency market capitalization (November 12, 2016) [1]. The project is open-source and financed by donations and the seven core developers [35].

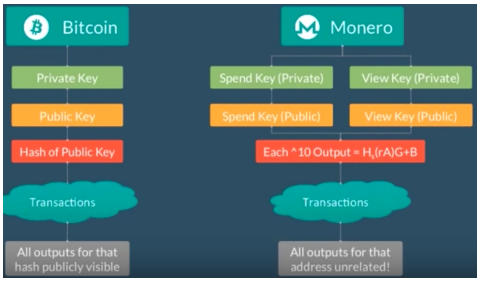
3.3.1.1 Description of the System

To achieve privacy, the CryptoNote technology behind Monero has features that are not implemented in classical Blockchain networks such as Bitcoin. In particular, there are two conditions that need to be satisfied for any currency to be private. Transactions need to be untraceable and unlinkable [36]. The conditions are defined below and the technical implementation in Monero is explained briefly. In addition, the other concepts behind Monero are introduced.

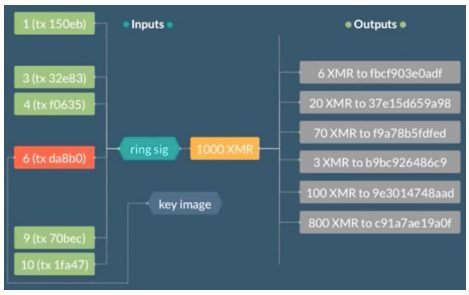
Definition unlinkable transactions: "For any two outgoing transactions it is impossible to prove they were sent to the same person."[36]

Definition untraceable transactions: "For each incoming transaction all possible senders are equiprobable."[36]

Unlinkability and Untraceability in Monero To achieve unlinkable transactions, Monero introduced a concept that is using two kind of keys. Instead of just a private and a public key, you have a spend and a view key in Monero. Both keys exist as private and public [38]. An overview of the keys and their context in comparison to Bitcoin is shown in fig. 3.7a.
What makes transactions unlinkable is the hash function, which takes the public view and spend key, as well as a random number as an input. Because of the random number, all outputs seem to go to different persons, even if they are sent to the exact same person. In addition to the randomized hash function, all inputs of transactions are split into multiple outputs as shown in fig. 3.7b on the right side [38]. This example presents a transaction of 123 Monero. They are split into 100XMR, 20XMR and 3XMR. Three more outputs are created to match the next power of ten as total amount of inputs and are just sent back to the sender. For an observer of the blockchain, it is now impossible to tell how much Monero were sent to which recipient [38].
To explain how transactions are made untraceable, the example mentioned above (shown in fig. 3.7b) is considered again. Monero uses a ring signature to sign transactions, which basically just tells the observer that one member of a certain group of people is the sender of the transaction. To always find enough members for the ring signature, the amount of an input always needs to be a power of ten. To prevent double spending, a key image, which is similar to a hash of the senders address, is added to the ring signature [38].



(a) Keys in Bitcoin & Monero [38]



(b) Transaction in Monero [38]

Figure 3.7: Overview of Monero's keys and transactions

Other Features of Monero In Monero the emission curve is smooth. Which means that the block reward is adjusted after every single block and not after several 10000 as in Bitcoin. Also the minimal Block reward is fixed at 0.3XMR, making the currency slightly inflationary with no limit of coins [36]. The proof of work algorithm used in Monero is memory intensive, making the design of ASIC devices harder. As this implies that Monero can be mined with normal PC's, the network should generally be more decentralized. For improved scalability, the block size is adaptive [36].

3.3.1.2 Performance

The amount of transactions per second that is possible is only limited by the hardware of the nodes. This is a consequence of the dynamic block size. A the moment 1700 tps are estimated, assuming a node using an an Intel i7 2600k quad core processor. The limiting factors of tps today are mainly bandwidth and memory [39].

3.3.1.3 Attacks and Challenges

The possible attacks in Monero are very similar to those in Bitcoin because they are both based on blockchain technology. However in Monero an attacker that is spamming the

network with many small transactions might have a bigger effect on the network because of the dynamic block size. As the blocksize grows the network gets overloaded.

3.3.2 Rootstock

Rootstock, or RSK abbreviated, is a distributed smart-contract platform similar to Ethereum and currently under development by the Argentina based RSK Labs. The RSK platform has three main features. Firstly it features a Turing-complete virtual machine that supports smart contracts. Secondly, it is a Bitcoin sidechain. And lastly RSK supports merged-mining through a federated consensus protocol. In addition RSK should provide other enhancements such as faster transactions and better scalability [16].

3.3.2.1 Description of the System

As the Rootstock platform is not released to the public yet, this section is mainly based on the Rootstock whitepaper and the developers statements. It summarizes the technology as well as some of the possible use cases for Rootstock as a Bitcoin Sidechain.

Smart Contracts Rootstock features an independent virtual machine, that is compatible with the Ethereum virtual machine at opcode level, thus the smart contracts running on Ethereum can also be run on RSK. That means that Rootstock provides Ethereum users the possibility to run their projects on top of the Bitcoin network [16].

Sidechain Rootstock works as a Sidechain to Bitcoin, meaning that Bitcoins can be transferred into the Rootstock Blockchain and become so called Rootcoins. These Rootcoins can also be transferred back into the Bitcoin Blockchain at any time and with no additional costs. There is no currency issuance, that means that all Rootcoins are created directly from Bitcoins [16].
The basic sidechain mechanism described above is oversimplified. In practice, the transfer between the two Blockchains is actually not a real transfer, but achieved by locking a certain amount of coins in one of the Blockchains and unlocking the exact same amount in the other Blockchain. Locking and unlocking instead of transferring coins is necessary, because the two Blockchains cannot verify the authenticity of the balances in the other Blockchain [17]. For this process to be fully trusted and third-party-free in both directions, smart contracts that specify on the locking and unlocking mechanism are needed on both platforms. Since Bitcoin does not support smart contracts, this two-way pegging system requires trust on a group of semi-trusted third-parties which execute the transfer [16]. These third-parties cannot control the locking and unlocking of coins on their own, but they can as a group. For Rootstock a federation of leading Bitcoin stakeholders such as exchanges, wallets and payment processors were chosen as the third-parties, as their incentives are aligned with the semi-trusted third-parties [16].

Merged-Mining Merged-Mining is the process of mining two cryptocurrencies at the same time with the same algorithm. Rootstock can be merge-mined with Bitcoin [16]. The fact that no additional hashing power is needed, gives Bitcoin miners an incentive to also mine Rootstock, even though they only get transaction fees and no block rewards. These fees then can easily be transferred back to Bitcoin [16].

3.3.2.2 Performance

Rootstock will be able to handle up to 300 transactions per second (tps) at its launch [**?**]. In comparison the Bitcoin network is restricted to a sustained transaction rate of

7tps. This is caused by the Bitcoin block size being restricted to 1MB per block and the slow block interval of ten minutes. Without a hard fork changing the block size, it will be difficult to increase the transaction rate of the Bitcoin network [2]. Even though RSK is scalable up to 1000tps without a hard fork, the problem of further scalability is still existing [16]. VISA handles around 2000tps on average today and its peak rate is at about 4000tps. Looking into the future, more payments will be made online, as a result the transaction rate will have to be even higher. If RSK or any other Blockchain was be able to handle all economic transactions, including cash, the transaction rate probably needed to be in the hundred thousands [2].
Bitcoin is a peer-to-peer network and nodes randomly connect to other nodes. As the network grows, the latency between nodes increases. To decrease the latency, some of the big Bitcoin miners developed a fast relay network [16]. This network consists of globally strategically placed nodes that miners can connect to, to decrease distance to the node and thus latency. Such a fast relay network is implemented in RSK from the beginning, reducing latency for miners but sacrificing total decentralization [16].

3.3.2.3 Attacks and Challenges

Merged-Mining comes with the risk of a possible 51% attack of one of the big Bitcoin miners. At launch of Rootstock, few miners that mine Bitcoin will also mine RSK, but those that do, will use their full hashing power. The probability for one of them gaining more than 50% of the entire hashing power of the RSK network is pretty high. As a consequence said miner is able to double spend [16].
If some of the members of the federation that supervise the transfer of coins between the Blockchains cooperate, they could be able to lock the coins on one Blockchain and not unlock them on the other. This event is unlikely to happen, because the members of the federation are highly respected Bitcoin stakeholders [16].

3.3.3 IOTA

The number of connected Internet of thing (IoT) devices is estimated to grow to 50 billion in the next decade [21]. Of course, for the realization of this huge network of devices many obstacles have to be overcome. One of the obstacles is, that all devices must be able to automatically and seamlessly pay tiny amounts to each other. IOTA is a cryptocurrency particularly trying to solve this problem, called micro-transactions. The project started in 2015 and is currently in beta [22].

3.3.3.1 Description of the System

Classic Blockchains such as Bitcoin have drawbacks though, especially concerning micro-transactions. Very small transactions make no economic sense due to constant transaction costs for example. IOTA tries to solve these drawbacks with a new blockchainless approach [18].

The Tangle Basically IOTA uses a directed acyclic graph (DAG) as the database instead of the Blockchain. This graph is called the Tangle of which an example is presented in fig. 3.8. Every time a new transaction is issued by a node of the network, a site is added to the set of the graph. Nodes of the Tangle graph are called sites, so that they cannot be mistaken for network nodes [18]. The node that issued the transaction now has to approve two existing transactions. In the DAG these approvals are represented by edges between two sites. The edges are directed from the approving site to the approval

site. The transaction A indirectly approves B, if there is no direct edge between them, but a path of at least length two [18].

Token Distribution At launch of the network, there is one single node holding all tokens. The very first transaction, the "genesis", distributes these tokens to several other nodes. After that, no additional tokens are issued. As a consequence, mining does not exist in IOTA (in the sense of nodes getting token rewards for approving transactions) [18].

The System in Action To issue a transaction a node has to complete three steps. Firstly, it uses an algorithm to choose two other transactions for approval. Secondly, it checks whether these two transactions are conflicting. If so, it does not approve them. Lastly, it solves a cryptographic puzzle similar to the Bitcoin proof of work [18].
If two transactions are conflicting with each other, they stay in the Tangle, but sooner or later, more of the new transactions will approve the correct one and the others will be orphaned [18].

Weight Transactions stored in the ledger have a weight that is proportional to the work that the node puts in to issue said transaction. In practice those values are always of the form 3^n. The cumulative weight of a transaction is defined as the sum of all weights of transactions approving this transaction, directly or indirectly, and the own weight [18]. In fig. 3.8, the weights are represented by the bottom right number and the cumulative weights are represented by the top left number. As one can see, newer transactions are always placed on the right side of older transactions in the graph [18].

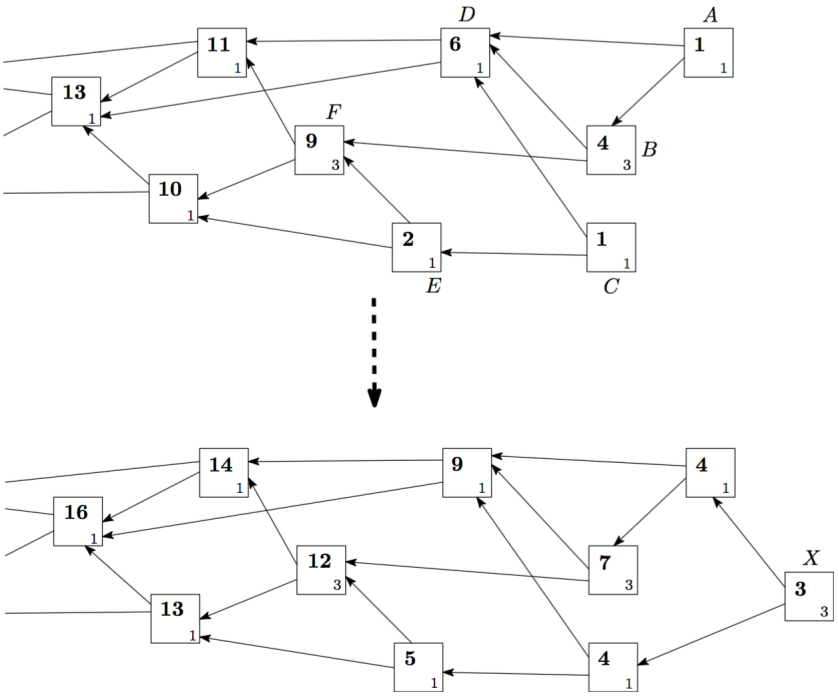


Figure 3.8: Weight recalculation after a new transaction [18]

As an example we look at the graph shown in fig. 3.8. Firstly, the transactions A and C are the only unconfirmed ones. As an example the cumulative weight of site F is considered. Before a new transaction X arrives the cumulative weight is defined as described above: A + B + C + E indirectly approve F, adding the sites own weight we get 1 + 3 + 1 + 1 + 3 = 9. Now a new transaction X with weight 3 is issued. It approves all before unconfirmed

transactions, thus the cumulative weight of all transactions increases by the weight of X [18].

3.3.3.2 Performance

Because of the totally different approach of IOTA and the fact that the system is still in beta, little theoretical or real performance data exists. However, one of the core developers stated that he reached 50 transactions per second in september 2016 [23].

3.3.3.3 Attacks and Challenges

The IOTA system is prone to attacks similar to classic Blockchains [18]. One of the attack types is when the attacker tries to outpace the system to double-spend. To do so in IOTA, the attacker could first spend his assets and later try to double-spend the same assets again, but this time giving the transaction a very big weight. The double spending transaction needs to outweigh the subtangle of the legitimate transactions and approve transactions not approving the legitimate transaction. This strategy leads to the legitimate subtangle being orphaned, but the product bought with the legitimate transaction might have already been shipped [18]. An illustration of such a large weight attack can be seen in fig. 3.9. This threat is real, if the weight of a single transaction can be infinitely large. To reduce the chance of a large weight attack, there needs to be a weight limit on transactions [18].

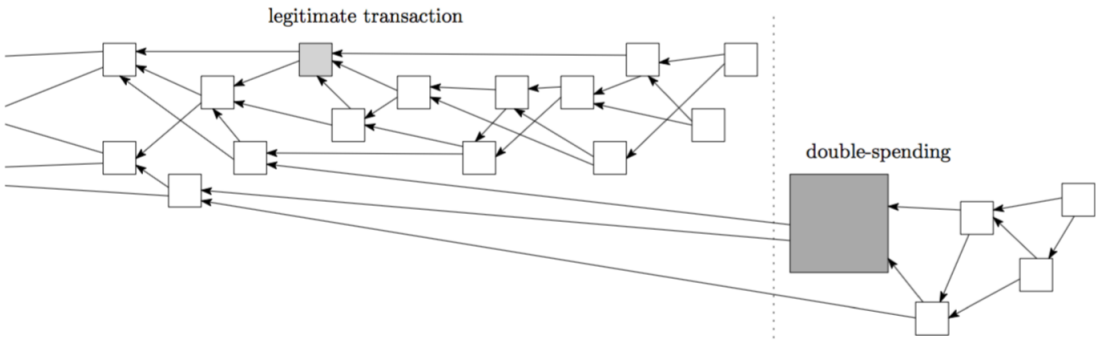


Figure 3.9: Large weight attack [18]

3.3.4 BigchainDB

The project around BigchainDB was started in the summer of 2015 by a small team around Trent McConaghy [24]. After working on ascribe.io, a bitcoin blockchain-based intellectual property attribution, and having to turn down potential customers due to scalability, they quickly realized the need of bigger throughput on blockchain technologies. Differently from Bitcoin with its only seven transactions per second limit, BigchainDB is supposed to handle one million transactions per second, scaling up the throughput considerably [2, 24].

3.3.4.1 Description of the System

BigchainDB is based on a RethinkDB cluster and adds blockchain features like immutability, decentralized control and the ability to create and transfer assets. RethinkDB is a scalable, decentralized open-source JSON database [25]. Each added node in a RethinkDB cluster brings its own storage to the pool, adding to the total storage of the cluster, as

shown in fig. 3.10. With this method storage pools of petabytes of data is easily achievable. This is to say without any replication of Data, any replication of the factor R would decrease the total storage capacity by a factor of R [26].

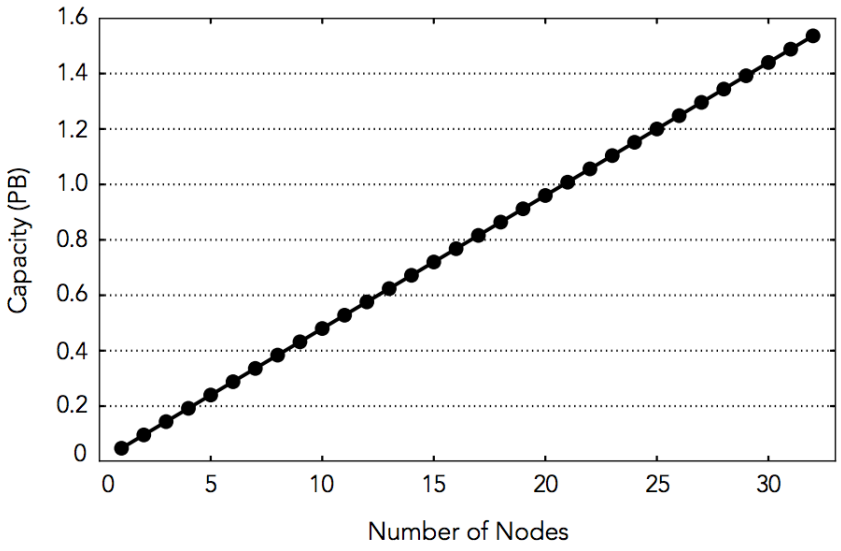


Figure 3.10: linear scalability of the storage size [26]

BigchainDB presents itself to an outside client as a single big blockchain, in reality however there are two different distributed databases inside of it. One database is the Backlog (S) and one is the actual Blockchain (C). These distributed databases run on different nodes who either have or do not have voting power, therefore creating a super peer-to-peer2network. A client submits its transactions to the Backlog first that consists of an unordered amount of transactions from different clients [26]. To be more precise, a client submits its transaction to a certain node, this node validates the authenticity of the transaction, according to that node alone and then submits it to the backlog. The receiving node also randomly assigns the transaction to another node, where it is stored later. The node that is running the BigchainDB Consensus Algorithm orders its list of transactions, creates a block of them and puts this block into the second distributed database (C) which is the actual blockchain [26]. This is shown in fig. 3.11. A node capable of signing can then vote whether this block is valid or invalid. When a majority vote is achieved the block goes from undecided to decided_vlid or decided_invlid. On the blockchain each block contains a hash of itself which is also its primary key. Each vote contains the hash of the previous block, therefore creating a chain. All nodes sign the blocks that they create with their private key. At this point, according to the BigchainDB developers, immutability is achieved [26].
All in all, BigchainDB offers a decentralized, immutable database that can be used to create and transfer assets. The developers propose the following use cases of BigchainDB: Tracking intellectual property assets, providing evidence in form of receipts and certification, storing legally binding contracts, creation and movement of financial assets and smart contracts [26].

3.3.4.2 Performance

BigchainDBs algorithm is designed to not be a limiting factor on its performance, meaning that the main limiter will be the underlying rethinkDB database. Performance experiments on rethinkDB show that with 32 nodes a throughput of one million transactions per second is achievable. The whole Database scales linearly with the amount of nodes in it. These tests, however, are done without any involvement of BigchainDBs algorithms

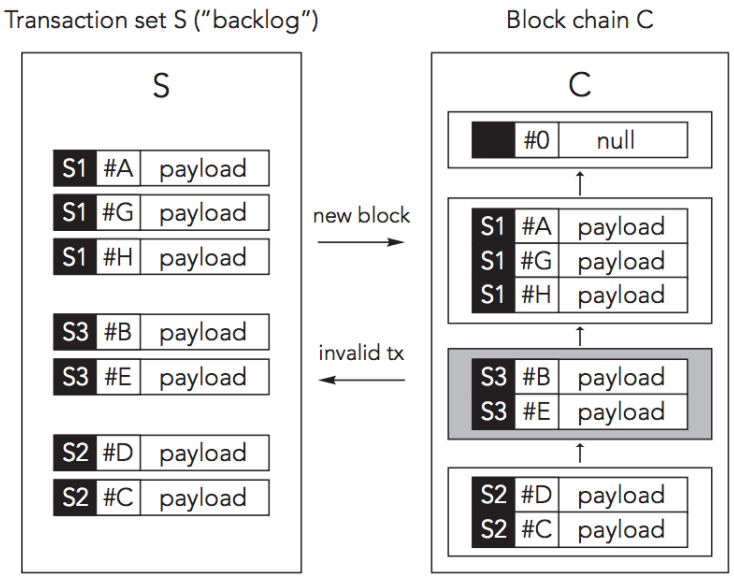


Figure 3.11: Transactions are formed to a block and transferred to the blockchain [26]

and future experiments need to be conducted to show whether their claims hold true or not [26].

3.3.4.3 Attacks and Challenges

BigchainDB is still in its early phase of development, therefore a lot of information is not available yet. Most of the standard blockchain attacks also apply to BigchainDB. Some attacks, like the sybil attack, are less of an issue here. A sybil attack creates multiple fake nodes to gain a majority vote in a system. Because of the private nature and cryptographic signed nodes of BigchainDB this should not be possible. On the other hand securing the data from malicious nodes is more of a problem since BigchainDB is built on a traditional Database, where a node with admin rights can simply drop all tables in its storage [27].

3.3.5 La'Zooz and Arcade City

La'Zooz as well as Arcade City are both projects that are trying to establish an owned car sharing platform. La'Zooz was established in 2014 in Israel. Independently began Arcade City in 2015 in the US. While La'Zooz focuses on its core as a ridesharing service, Arcade City has planned a broader application including deliveries and short term home rentals such as Airbnb. Both of these services are still in early development [28, 32].

3.3.5.1 Description of the System

La'Zooz tries to use currently unused capacities in today's transportation network. The basic idea is that someone can request a ride from A to B and the network automatically informs a nearby driver, that is already heading in this direction, to pick the person up. Upon completing this ride the driver will receive a certain amount of road zooz, La'Zoozs own cryptocurrency, from the rider. Road zooz's coins can either be mined or purchased and then used to catch rides. It is possible to only buy road zooz and never drive on your own. The system encourages sharing your own vehicle and therefore increases the transport density [29].
Unlike La'Zooz who deploys its own currency, Arcade City is built on the basis of Ethereum where it creates its own currency that is called ARC. A fixed amount of ARC tokens will be released in form of a token sale and other means, as shown in fig. 3.12 [32].

Figure 3.12: Distribution of ARC Tokens [32]

Other payment methods are also planned but the ARC token will be incentivized with less fees. The process of catching a ride is done like this: A rider puts in start and end location of his desired ride and is then shown the price. This is broadcasted to the network via an Ethereum whisper, for anyone to see [32]. A driver can then claim this drive. Upon doing so both the driver and the rider pay the cost of the ride into a pool. At the end of the ride the driver will receive the full amount minus the Arcade City fees. This is done via a smart contract. If there would be a dispute the user could refuse to release the fund and start resolving the issue [32].
In order to compete with already established ridesharing services Arcade City as well as La'Zooz both implement a small game, in order to gain a big enough user base in a city [34, 30].

3.3.5.2 Performance

Since Arcade Cities ARC tokens are based on Ethereum, the performance can be equaled with it. This means 30 transactions per second at the moment. Since Arcade City plans to use other forms of payments this would lead to an increase in their performance, although it is likely to have trade offs in its blockchain features [32]. Of Lazooz performance is not much known.

3.3.5.3 Attacks and Challenges

Attack vectors are similar to other blockchain applications such as the sybil or double spending attack. However, since Arcade City builds on Ethereum it should have a strong basis against such attacks. A bigger problem with these technologies is funding and trust. Since the beginning of October 2016 La'Zooz's servers are no longer in operation. This is presumably because they failed to reach their target in a crowdsale. Further information is not available at the moment [31].
Arcade City fights with different accusations, some even going as far as calling them a complete scam. Their released app did not contain any true ridesharing functionality and was seen as a move to stall for time [33].

3.4 Comparing blockchains

Comparing all these blockchain technologies is not entirely possible. They are too different in a a lot of aspects to make an effective comparison. Therefore it can also be viewed as an overview of the features of these technologies, with only certain aspects being drawn to a good comparison.
Under "Performance" criterias like transactions per seconds can be decently compared across all technologies and general criterias like the size of the community, though diffi-

cult to measure, are also good comparing points. Comparing their currencies or smart contracts is a lot more subjective, however, some do not even have these features. Where it was possible we compared actual numbers, on criteria where this was not possible we restrained ourselves to simply mention whether a certain feature is available or not. Most of the following statements reference our collected information above from each of the respective technology.

3.4.1 Comparison Criteria

Our comparison criteria can be roughly divided into three groups. Project scope, performance and production readiness. In the first group, project scope, we present an overview of what this technology composites of.

3.4.1.1 Project Scope

All of our explored blockchains besides BigchainDB offer some form of a currency. Some create their own, others simply build on top of already existing ones. Rootstocks premise is that it brings smart contracts onto the Bitcoin blockchain and therefore does not have a real currency on its own but will use Bitcoins themselves. Rootstock as well as Ethereum offer fully functional real smart contracts. BigChainDB has only a limited control over smart contracts and can run a subset of smart contracts but not arbitrary contracts [40]. La‘Zooz as well as Arcade City use smart contracts to ensure payment for their rides, they do not offer arbitrary code execution tough.

3.4.1.2 Performance

From the performance criteria important aspects are the number of possible transactions per second as well as confirmation time. Since all of these relatively new technologies are still in development we added a third point of scalability which expresses the capabilities of scaling their technologies, either through measurements that are already in place or in the future planned developments.
The number of transactions per second is relatively straightforward. Some exceptions however are Monero and IOTA where their limit is not a possible to be calculated in a precise manner but is more dependent on the hardware used. The values we chose are generally accepted estimates at the time of writing this paper [39, 18]. BigchainDBs impressive 1000000 transactions per second was achieved through an experiment by the developers without the full functionality of their product yet [26]. Since the Arc tokens of Arcade City are built on Ethereum a similar performance can be expected with them. Since they accept different payment methods than Arc tokens a reliable number can not be said however. No estimates for La‘Zoozs tps as well as confirmation time are known at the moment.
Ethereum, Rootstock as well as Monero have concrete plans already on how to increase their performance in the future. IOTA as well as BigchainDB are already scalable by their current design. Bitcoin has a number of different propositions on solving its scalability issues, there does not seem to be a consensus on what to do yet tough.

3.4.1.3 Production Readiness

Bitcoin, Ethereum as well as Monero have released fully functional products already. Rootstock has not yet shown anything and IOTA is in a closed Beta at the moment. BigChainDB has a proof of concept client released that is not yet production ready but shows their current progress [41]. Arcade City as well as La‘Zooz have both released early versions of their clients through the google play store. La‘Zoozs client is currently not

functional however and Arcade Cities app is merely a placeholder without any serious functionality. Bitcoin and Ethereum have a large community around them, both developers as well as enthusiasts. The newer technologies still have some catching up to do here.

3.4.2 Comparison

According to the criteria introduced and explained in Section 3.4.1 and the detailed explanation of each technology in Section 3.2 and 3.3, we have created an overview over all introduced blockchain services.

Table 3.1: Comparison of different Blockchain services

	Bitcoin	**Ethereum**	**Rootstock**	**Monero**	**IOTA**	**BigchainDB**	**La'Zooz / Arcadecity**
Currency	✓	✓	(✓)	✓	✓	×	✓
Smart Contracts	×	✓	✓	×	×	(✓)	(✓)
Public Chain	✓	✓	✓	✓	✓	✓	✓
Private Chain	×	✓	×	×	×	✓	×
Transactions per Second	3 - 4	30	300	1700	100	1000000	unknown
Immutability	60 min	3 min	30 sec	12 min	˜1 sec	˜1 sec	unknown
Scalability	+ −	+	+	+	+ +	+ +	+ +
Established User Base	+ +	+	− −	+	+ −	−	−
Released	✓	✓	×	✓	Beta	Poof of Concept	Discontinued / Placeholder App
Community	+ +	+ +	−	+	+ −	− −	−
Future	+	+	+ −	+	+ −	+ −	−

3.5 Summary and Conclusion

To be able to compare different blockchains, this paper first explains the technology and its history on the basis of Bitcoin. Further concepts important for the comparison are discussed by means of Ethereum. Then, six different blockchain services are presented: Monero, Rootstock, IOTA, BigchainDB, Arcade City and La'zooz. The choice of which services to compare was not an easy one, because there are so many arising platforms and the blockchain market is very young and still very unstable. The services to compare are chosen to show different use-cases of blockchain technology. As a consequence some services are in a niche market and might already have been discontinued.
Because of the totally different approaches and the fact that there is missing or unconfirmed data for some projects, an exact comparison of all aspects is not possible. The comparison criteria are defined vaguely and grouped into three categories: project scope, performance and production readiness. The services are then compared and their differences as well as their strengths and weaknesses are shown. As a result of the comparison, the collected data is broken down into a single compressed table. The blockchain technology in general looks promising, but to establish the technology in the future to use it as a distributed database to save any kind of data, it needs more improvement.

Bibliography

[1] Crypto-Currency Market Capitalizations; https://coinmarketcap.com, November, 2016.

[2] Scalability; https://en.bitcoin.it/wiki/Scalability, November, 2016.

[3] Satoshi Nakamoto: *Bitcoin: A Peer-to-Peer Electronic Cash System*; White Paper, 2008. https://bitcoin.org/bitcoin.pdf

[4] Michael Crosby, Nachiappan, Pradhan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman: *BlockChain Technology*, Sutardja Center for Entrepreneurship & Technology, Berkeley Engineering, University of California, October 2015. https://pdfs.semanticscholar.org/4b65/d3eda63fc18303dfbc071fece0e276a7a16c.pdf

[5] Ken Shirriff's blog; http://www.righto.com/2014/02/bitcoins-hard-way-using-raw-bitcoin.html, November, 2016.

[6] Björn Segendorf: *What is Bitcoin?*; Sveriges Riksbank Economic Review, Stockholm Sweden, 2014. http://www.riksbank.se/Documents/Rapporter/POV/2014/2014_2/rap_pov_artikel_4_1400918_eng.pdf

[7] Bitcoin Block Reward Halving Countdown; http://www.bitcoinblockhalf.com, November, 2016.

[8] Bitcoin Forum, Topic: 7 transaction per second limit still true?; https://bitcointalk.org/index.php?topic=1391143.0, November, 2016.

[9] What can an attacker with 51% of hash power do?; http://bitcoin.stackexchange.com/questions/658/what-can-an-attacker-with-51-of-hash-power-do, November, 2016.

[10] Moore's Law or how overall processing power for computers will double every two years; http://www.mooreslaw.org, November, 2016.

[11] Vitalik Buterin: *Ethereum Whitepaper A Next-Generation Smart Contract and Decentralized Application Platform*; White Paper, 2013. https://github.com/ethereum/wiki/wiki/White-Paper

[12] Ethereum Community *Ethereum Homestead Documentation*, Release 0.1, October 2016. http://www.ethdocs.org/en/latest

[13] Formalizing and Securing Relationships on Public Networks by Nick Szabo; http://ojphi.org/ojs/index.php/fm/article/view/548/469, November 2016.

[14] Introduction to Smart Contracts; http://solidity.readthedocs.io/en/develop/introduction-to-smart-contracts.html, November, 2016.

[15] What number of confirmations is considered secure in Ethereum?; http://ethereum.stackexchange.com/questions/319/what-number-of-confirmations-is-considered-secure-in-ethereum, November, 2016.

[16] Sergio Demian Lerner: *RSK Bitcoin powered Smart Contracts*; White Paper, Revision 9, November 2015. http://www.the-blockchain.com/docs/Rootstock-WhitePaper-Overview.pdf

[17] Sergio Demian Lerner: *DRIVECHAINS, SIDECHAINS AND HYBRID 2-WAY PEG DESIGNS*; RSK LABS LTD., Revision 9, April 2016. http://www.the-blockchain.com/docs/Drivechains\%20sidechains\%20and\%20hybrid\%202-way\%20peg\%20designs\%20-\%20Sergio\%20Lerner\%20-\%202016.pdf

[18] Serguei Popov: *The Tangle*; White Paper, Version 0.6, Jinn Labs, April 2016. http://iotatoken.com/IOTA_Whitepaper.pdf

[19] What is IOTA?; https://iota.readme.io/v1.1.0/docs, November, 2016.

[20] Bitcoin Forum, Topic: IOTA; https://bitcointalk.org/index.php?topic=1216479.0, November, 2016.

[21] By 2025, Internet of things applications could have $11 trillion impact; http://www.mckinsey.com/mgi/overview/in-the-news/by-2025-internet-of-things-applications-could-have-11-trillion-impact, November, 2016.

[22] The Anniversary Compendium; https://medium.com/iotatangle/the-anniversary-compendium-18cd74d6abd3#.q63jcn834, November, 2016.

[23] Currently seeing 50+ TPS in #IOTA. Going to test 100+ TPS this weekend. #beyondblockchain #tangle; https://twitter.com/DomSchiener/status/776483482814058496?lang=de, November, 2016.

[24] BigchainDB: Where it Came From, Where We‘re At, Where We‘re Headed; https://blog.bigchaindb.com/bigchaindb-where-it-came-from-where-were-at-where-we-re-headed-5004a319e35f#.3mlmt8n2i, November, 2016.

[25] Frequently asked questions: What is RethinkDB?; https://www.rethinkdb.com/faq/, November, 2016.

[26] Trent McConaghy, Rodolphe Marques, Andreas Mueller, Dimitri De Jonghe, Troy McConaghy, Greg McMullen, Ryan Henderson, Sylvain Bellemare, Alberto Granzotto: *BigchainDB: A Scalable Blockchain Database*; White Paper, ascribe GmbH, Berlin, Germany, June, 2016. https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf

[27] How BigchainDB is Decentralized; https://docs.bigchaindb.com/en/latest/decentralized.html, November, 2016.

[28] La‘Zooz ABOUT US; http://lazooz.org/index.html#about, November, 2016.

[29] members of the La‘Zooz community: *La‘Zooz White Paper*; White Paper, June, 2015. http://www.the-blockchain.com/docs/LaZooz\%20Blockchain\%20Taxi\%20Whitepaper.pdf

[30] Google Play: La'Zooz; https://play.google.com/store/apps/details?id=com.lazooz.lbm\&hl=en, November, 2016.

[31] Is La'Zooz dead?; https://www.reddit.com/r/LaZooz/comments/4v1279/is_lazooz_dead/, November, 2016.

[32] Christopher David, Stefaan Ponnet, Kristien De Wachter, Ben Adriaenssen, Michael Thuy: *Whitepaper and Token Plan*; White Paper, v 1.2, arcade.city, October, 2016. https://arcade.city/AC-whitepaper.pdf

[33] Proof: Arcade City is a Scam; https://medium.com/@CTUAgentIvan/arcade-city-is-a-scam-98c22c557f18#.rko7u969i, November, 2016.

[34] arcade.city; https://ac01.netlify.com/, November, 2016.

[35] ABOUT MONERO; https://getmonero.org/knowledge-base/about, November, 2016.

[36] Nicolas van Saberhagen: *CryptoNote v 2.0*; White Paper, October, 2017. https://cryptonote.org/whitepaper.pdf

[37] Surae Noether: *REVIEW OF CRYPTONOTE WHITE PAPER*; July, 2014. https://downloads.getmonero.org/whitepaper_review.pdf,

[38] How Monero Took Over Bitcoin's Unique User Base; https://www.youtube.com/watch?v=N5-Kqyr4BQI, November, 2016.

[39] How many transactions per second can the Monero network handle?; http://monero.stackexchange.com/questions/405/how-many-transactions-per-second-can-the-monero-network-handle, November, 2016.

[40] BigchainDB and Smart Contracts; https://docs.bigchaindb.com/en/latest/smart-contracts.html, November, 2016.

[41] Production-Ready?; https://docs.bigchaindb.com/en/latest/production-ready.html, November, 2016.

Chapter 4

Emerging Pricing Models for Cloud Services

Sebastian Elke, Laurenz Shi, Linda Samsinger

While cloud computing has received increasing attention due to its profitability in recent years, comparing pricing and price models of different cloud services providers has been little explored academically. Amazon, Google and Microsoft have started price races of cloud services, featuring infrastructure, platform and software as a service. We constructed a theoretical framework for comparing the impact of different pricing models and compiled prices from five different cloud service providers to determine underlying pricing patterns and to understand the price differences across different cloud service providers. Cloud providers adopt different pricing models, which can determine the economic success of cloud services. Competition among cloud service providers has driven growth and innovation of pricing models and the widespread availability of services to cloud users. We find that companies that require scalability of resources will favour dynamic pricing schemes, while smaller companies are more interested in the predictability of cloud service costs. Our study of different cloud service prices shows that Google consistently charges the lowest prices and MiroNet—a small-scale Swiss cloud service provider—charges the highest prices regardless of differences in cloud computing configurations.

Contents

4.1 Introduction and Problem Statement

Rapid development in cloud computing in recent years has given rise to a very profitable business sector worth investigating. Clouds such as Dropbox, Google Drive, the iCloud, Spotify, Youtube, iTunes and OneDrive are known to a wide public.

4.1.1 Introduction

The major function of a cloud computing system is storing data on the cloud and using technology on the client to access that data [29]. Several business models rapidly evolved to harness this technology by providing software applications, programming platforms, data-storage, computing infrastructure and hardware as services [29]. Many companies offer various services in cloud infrastructure for organizations across the globe. The cloud trend has grown in parallel to the steadier development of on-premise infrastructure and traditional computing. Thus, competition of prices in the global market has increased between on-premise server providers and cloud-based service providers and among cloud-based server providers as well. They are subject to the provider's goal of maximizing revenue by their price schemes. On the consumer's side, however, the main goal is to maximize services and functions for the lowest price [12].

4.1.2 Problem statement

Purchasing for Cloud Services is not easy, given the multitude of services, prices and Cloud Service Providers that exist on the internet. To shed some light on this problem, this paper will attempt to give answer to the question of which Cloud Provider is most affordable. Pricing models set up by Cloud Providers want to charge Cloud User with the maximum price, but are challenged by market competition factors and Cloud Users, who seek out the lowest price in a multi-provider market with the maximum quality of Service QoS. Since the goal is to figure out the lowest price with the maximum amount of services provided by the Cloud, the customer's side and perspective on prices is taken. This leaves out any analysis on the Cloud Provider's behalf for fixing costs that need to be covered by the price offered and building a corresponding cost model.
The Cloud Price Index [31] reveals that cloud buyers who mix services such as compute and storage from multiple providers by shopping around for the cheapest services for each part of the Web application can make substantial savings. Cloud buyers mix and match services, make long-term commitments, negotiate and volume discounts can make up to 58 % savings for a small application and 74 % for a large application. The Cloud Price Index specifies services required to operate a typical Web server application including compute, storage, databases, management etc., but is not available outside of North America and for private clouds. However, the complexity with dealing with latency between datacenters, managing different GUIs and APIs, invoicing, documentation, and support functions leaves most users paying a premium for an integrated solution form a single provider [31]. This study will examine, which Cloud Providers offer the cheapest price and rule out the fact, that customers are able to shop around for individual features at different Cloud Providers in accordance with their specific needs. Looking at a package of services that are offered by Cloud Providers makes it easier to compare prices among Cloud Providers. Also, customized solutions and a definition of a package of features needed from the customer's side must be made.

4.2 Definitions and theoretical framework

4.2.1 Fundamentals

Cloud computing technology requires some basic knowledge about the underlying and contributing components. These fundamentals will be discussed in the following sections.

4.2.1.1 Main components of a computer

The **von Neumann architecture** is a model for the structure of a computer by the physicist and mathematican John von Neumann et al. of 1945 [18]. Although the von Neumann architecture is an older model, it is still broadly applicable to the architecture of modern computers. As shown in figure 4.1 the von Neumann architecture is composed of a processor, the main memory, system bus, and I/O units [19]. The **central processing**

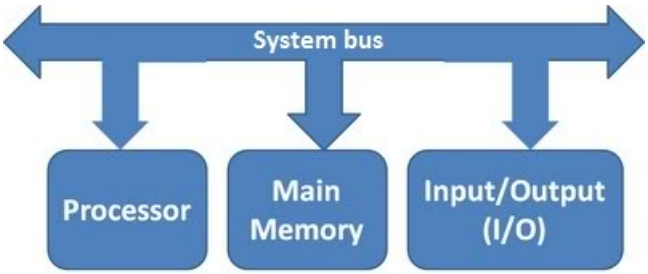


Figure 4.1: Von Neumann architecture [20]

unit, abbreviated **CPU**, is capable of reading and executing arithmetic and logic operations on programs. In addition, it controls and manages the hardware of a computer. These functionalities are core to every computer [19]. **Random access memory**, abbreviated **RAM**, serves as main memory of a computer. RAM saves programs and the data belonging to a program. It is possible to write into RAM as well as to read from it and as the name suggests, it is possible to access every byte in the memory directly. Therefore, it has few overtime and is fast [19]. **System buses** connect and serve as a communication link between the different parts of a computer [19]. The data to be processed and the processed data gets transferred via **input and output units** (**Input/Output**) from or to the environment [19].

4.2.1.2 Virtualization

A **virtual machine**, abbreviated **VM**, is a term that is used widely. Regarding the use of virtual machines in cloud computing, it describes the virtual duplicate of the real computer hardware on which software can run [23]. This allows to run programs independently of others, which can be used as far as running multiple operating systems on the same hardware, which is also referred to as **full virtualization**. This makes it possible for a single computer to provide the functionalities of multiple computers [22], which in the context of cloud computing is very beneficial due to the different environments and softwares in use. The **physical CPU** is the CPU of the real computer hardware. The **virtual CPU**, abbreviated **vCPU**, is the virtual computing power a VM gets provided for its own computations [23]. **Virtual RAM**, abbreviated **vRAM**, is, in the context of this report, the virtual memory assigned to the VM.

4.2.1.3 Service

Service is a broadly used term. It is most known in economics, where it describes an intangible product ready for sale. Service products can have many different characteristics, but in the economy and cloud computing field, two appear generally. Services are intangible and they occur mostly more than once [25]. Cloud computing is an ongoing business relation between the providers and users and therefore, quality and reliability of the services become an important aspect. The demands of users can differentiate heavily, which makes it hard for providers to satisfy them all. As a result providers negotiate with the users to find an agreement for the services provided. This agreement is called a **service level agreement** [24]. **Quality of service**, abbreviated **QoS**, is an important part of software and its pendant. Due to many different product features and user requirements, the quality of a software can be assessed by a grouping its characteristics. Accurate and comprehensive measurements of QoS can be helpful to illustrate values and advantages of a product to customers. The ISO/IEC 25010 suggests eight quality attributes for system and software products:

1. Functional suitability
2. Performance efficiency
3. Compatibility
4. Usability
5. Reliability
6. Security
7. Maintainability

These attributes should cover a wide range of functions to be able to satisfy the variety of users of system and software products [26]. QoS attributes generally are in the SLA and need to be enforced by monitoring those [24].

4.2.1.4 On-Premise Software

On-premise software is software, which gets installed on machines that are under the responsibility of the consumer. This stands in contrast to cloud computing where the provider installs the software on its own machines and the user accesses them over a network [27].

4.2.2 Cloud Computing

The following section is concerned with explaining the characteristics of cloud computing and its uses and should provide the information for the comparison of cloud service providers.

4.2.2.1 What is Cloud Computing?

Cloud Computing can be defined by what it is not. Before Cloud Computing came into existence, traditional computing or ”on-premise” computing was common place. Hardware had to be bought or rented at a local shop or online and software needed to be installed. The computer user was responsible for hard- and software maintenance. Cloud Computing can be accessed via the Internet and its scalability is what sets it apart from traditional Computing. Many authors have tried to define Cloud Computing such as ”It is driven

by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet [28]." It is the deliverance of computing as a service rather than a product, wherein shared resources, software, and information are provided to computers and other devices as a metered service over a network [14]. Cloud computing is an emerging practice for the online provisioning of computing resources as services [9]. The most analytical definition was made by Mell: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models [17]."

4.2.2.2 Essential Characteristics of Cloud Computing

Essential Characteristics [17]:

- **On-demand self-service**. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- **Broad network access**. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- **Resource pooling**. The providers computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.
- **High elasticity**. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
- **Measured service**. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

The characteristics of Cloud Computing are essential to the nature of Cloud Computing. The three service models and four deployment models are defined under "approaches", because they are part of the criteria customers use to determine what Cloud Services they should buy.

4.2.2.3 Advantages Cloud Computing

- **Lower initial investment costs**: customers do not need to pay CAPEX up front purchases of hardware or software anymore. No or a few servers are needed in the office. Cloud Computing services augment - and even replace - onsite infrastructure. This saves on equipment, management and data center floor space, and enables

organizations to gain greater business agility and flexibility [11]. Developers with inventive ideas for new Internet services no longer require the large funds in hardware to set up their service or the human expense to operate it [14].

- **Outsourcing of maintenance**: installation and update of software and maintenance of hardware are passed on to the Cloud Service Providers. Customers do not need to look at maintenance, backup and management costs and time.

- **Scalability**: endless potential of expansion, scalable according to requirements. Cloud Computing can alleviate resource poverty encountered with CPU performance, provide data storage capacity and processing power (Amazon Simple Storage Service S3, Flickr, Facebook) and divide application services effectively (optimal partition of application services, low network latency, high network bandwidth, adaptive monitoring of network conditions) [29] Cloud computing allows scalable on-demand sharing of resources and costs among a large number of end users [9].

- **Access independent of location**: Cloud can be accessed from everywhere. Flexibility of work location and hours ensues. Cloud computing customers can access their data wherever they are via the Internet [12]. Customers access their data worldwide as long as an Internet connection is available [9]. Cloud computing services are provided on 24/7 basis - anytime, anywhere [14]

- **Security experiences of provider**: cloud Computing overcomes obstacles related to security (e.g. reliability and privacy) [29].

- **Compatibility of OS**: Cloud Computing promotes flexibility and incessant scalability of IT resources that are presented to end users as a service through the internet medium. With a click of button operating systems can be switched from Linux to Windows within the same browser. On-premise infrastructure are dependent on pre-installed operating systems and the switching is relatively more complicated.

4.2.2.4 Disadvantages Cloud Computing

- **Cloud computing is challenged by an absence of standards**: limited scalability, unreliable availability of a service, service provider lock-in (absence of portability), unable to deploy service over multiple Cloud computing Service Providers CCSP (absence of interoperability) [29].

- **Network access dependent**: Cloud users need an internet connection to access the cloud environment. Dependency on the internet sharpens the digital divide the more dependencies are created over the internet.

- **Data security and regulations**: the single most problematic issue with cloud user is data security on clouds. As long as third parties can access and screen all data on a cloud, consumers are careful with the provisioning and transferring of confidential data to the cloud.

- **Cloud Service Provider in control**: the Cloud Service Provider is in control of the virtual hard- and software. Consumers cannot try to fix problems by themselves as they could with on-premise computing infrastructure. The intangible character of the cloud and the Cloud Service Providers makes it difficult for cloud consumers to trust in the cloud.

- **Singularity of software**: consumers are obliged to use the newest version of software for lack of an older version. Constant updating forces the consumer to adhere

to the newest software version without access to older software versions and use their operability in a specific context and positive features.

- **Control update policy**: the cloud consumer cannot access the cloud during control updates. If a cloud users needs to access the cloud urgently, the instance undergoing an update process will not be available.

4.3 Approaches

In this section, different approaches to evaluate Cloud Service Providers' pricing of cloud services are defined and discussed. The approach will be the method of analysis of cloud services providers according to different criteria.
A customer in want of a service of a cloud service provider will aim at the cheapest base plan price offered for the best quality and quantity of service. If the same services are offered by cloud providers, the customer will get the service, which is cheaper. Looking at equal services and thereby comparing apples with apples will make cloud providers comparable to each other.
The cloud services offered vary substantially across different providers. For the sake of comparison, similarity and equivalence must be established between Cloud Providers because comparing apples with oranges is impossible. In the following, all the different kinds of services will be listed, by which the customer forms a yardstick of reference.

4.3.1 Criteria

A customer will evaluate a prospective service provider based on tangible and intangible parameters [12]. While Cloud Service Providers shape pricing models according to their interests of revenue maximization, cloud customers' main goal is to obtain the highest level of service for a reasonable price. Cloud users will favor service provider offering based on the best QoS, fairness, pricing approach and utilization period with the lowest price [9]. The price charged should optimally satisfy both parties' interests. The Cloud Service Provider faces a trade-off: customer loyalty is compromised by a one-off charging of the highest price possible. The price must consider costs of the Cloud Service itself, prices offered by the market competition, and how the customer values the service or product offered in view of future cash flow streams [9]. Cloud customers can select different Cloud Services based on a variety of criteria that group Cloud Services into different categories. The most important criteria are listed below:

4.3.1.1 Service Models

The most common differentiation of Cloud Services is based on the type of service they provide. Starting at the lowest level:

Infrastructure as a Service It is the capability provided to the consumer to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls) [17]. Examples are Amazon Web Services EC2 and EC3.

Platform as a Service It is the capability provided to the consumer to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The

consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment [17]. Examples are Google App Engine and Azure App Service.

Software as a Service It is the capability provided to the consumer to use the providers applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings [17]. Examples are Facebook, GoogleDocs, YouTube and SAP Cloud Solutions.

Beyond these basic three service models, many more services such as Database as a Service (DBaaS), Storage as a Service (STaaS) by subscription, Security as a Service (SECaaS) integrated in a cooperative infrastructure and Test Environment as a Service (TEaaS) etc. exist.

4.3.1.2 Pricing Models

There are many different models in cloud computing. However, pricing models can be distinguished among pricing schemes and pricing mechanisms. Pricing schemes employed by Cloud Service Providers are: usage fee (pay-as-you-go hourly rate) scheme, subscription fee (monthly flat rate) and advertising as is the case with YouTube. These are highly competitive markets mainly in SaaS. The Cloud Service becomes free of charge for users and costs of the cloud service are covered by ad payments. A more complex model with similar charachteristics is called freemium. In a freemium model, there is a basic version which is for free and an integrated premium version that can be paid by the user to have access to the full-fledged program. The free version in a freemium model can either have advertisement or not.
Pricing mechanisms are divided into two pricing models: the fixed pricing and the dynamic pricing mechanism. Among the fixed (static) "menu" mechanisms, defined prices are based on static variables. In Cloud Computing Services this includes fixing list prices (Green.ch) and volume dependent prices, where price is a function of the quantity purchased (VMware). Static pricing models have fixed prices for pre-defined periods of time, making future payments highly predictable. Static models have a defined set of services that come with the booked package. A static or fixed pricing mechanism charges the customer the same amount of money at all times [9]. A static pricing scheme does not benefit the service provider because it does not reflect the current market value, unless the fixed prices come at a significant discount compared to the dynamic prices (AWS reserved instances vs GCP). Dynamic Pricing describes prices that change based on market conditions. Negotiation (SLA individual agreements) and auctions (AWS spot instances) are two of the dynamic pricing mechanisms employed in the Cloud Computing world. Within a market-dependent pricing scheme, the customer is charged based on the real-time market conditions such as bargaining, auctioning, demand behavior and yield management [9]. Dynamic models are less predictable in terms of prices as they depend on additional input parameters that are dependent on the effective usage of the services. In a dynamic or differential pricing scheme, the price charges change dynamically according to service features, customer characteristics, amount of purchased volumes, or customer preferences [9]. The dynamic pricing scheme changes as time changes. According to the demand of a resource, the pricing is done dynamically so as to maximize the profit of the service provider [14].

The Cloudscape is dominated by dynamic pricing mechanisms and only few static mechanisms.
The choice of which pricing models to implement is shaped by the interests of the cloud computing provider on the one hand and the customers of cloud computing providers on the other hand. End users will favor service provider offerings based on the best QoS, fairness, pricing approach and utilization period with the lowest price [9]. A cloud computing provider's typical goal is to maximize revenues with its employed pricing scheme, while the customers' main goal is to obtain the highest level of QoS feasible for a reasonable price. The price charged should optimally satisfy both parties interests. Service providers can encourage the usage of its services by controlling for this important metric. The service provider seeks customer loyalty and the achievement of higher revenues at the same time. The price must consider the manufacturing and maintenance costs, market competition, and how the customer values the service or product offered [9]. The utilization period needs to be determined as well, which can be defined as the period in which the customer has the right to utilize the provider services based on SLAs between the two parties. While prices are set by models, often market reality determines what actual prices are paid. Thus, highly competitive markets' prices change quite often and usually break down due to competition.
Pricing models on the market, have all kinds of exotic names. It can be quite confusing to determine what kind of cloud pricing model is being used. Thus, dynamic prices can vary between periods while in static models the price remains unchanged for all periods. However, there are also hybrid solutions. The more sophisticated a model, the less likely it is implemented.

4.3.1.3 Deployment Models

There are mostly three and sometimes four different deployment models utilized in the Cloudscape: the private cloud, the public cloud and the hybrid cloud [9]. The famous NIST definition of Cloud Computing appends the list by yet another deployement model—the community cloud—as an alternative [17].

1. In the **public cloud model**, the cloud resources and services are made available to the general public over the Internet. These services can be free or charged per use.

2. In the **private model**, the cloud computing environment is made available exclusively to a certain organization.

3. In the **hybrid cloud**, the service provider is in charge of external computing resources in addition to the internal resources. It can be viewed as composed of a public cloud and a private cloud. Its composition is made of two or more distinct cloud infrastructures (private, community or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability.

4. In the **community cloud**, the cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns. It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

4.3.1.4 Customer Segmentation

1. **Type** Customers can be differentiated between enterprise, private end users and start-up customers. Companies are split into big and small companies according to

size. Companies can be distinguished according to international or national business operations. Some customers need special treatment because of specific regulations that govern their type of business such as companies with classified and proprietary data.

2. **Industries** Cloud Service Providers specialize in different industries such as Banking & Finance, Government, Healthcare & Pharmaceuticals, Internet & Communications, Manufacturing & Engineering, Media & Entertainment, Retail & Customer Services.

4.3.1.5 Cloud Services

Depending on what the the customer needs on Cloud, different Cloud Services are offered and the products are given corresponding names. The broad variety of services makes price comparison difficult. Services range from Compute, Storage and Databases, Networking, Big Data, Machine Learning, Management Tools to Developer Tools, Identity I& Security services. Additional services include:

- **IP/VIP management**: manage network creation such as VLANs, Virtual IP addresses, and adding or removing hosts
- **messaging services**: manage messaging queues, pushing and receiving application messages from the cloud
- **storage services**: block storage, file storage, cloud storage, object storage, flexible storage services
- **server cloning**: create multiple copies of VM without re-installation and re-configuration
- **system monitoring**: real time statistics and analysis console to monitor cloud computing resource
- **VPN access**: access cloud infrastructure resources via a virtual private network VPN
- **control interface**: web-based application/control panel, graphical user interface

4.3.1.6 Quality of Service

The quality of service describes the requirements of services a Cloud Service Provider should offer to his customers. If the service provider ensures that these requirements are maintained at a high level, the quality of the service offered will increase. This will increase the number of customers and loyalty to the service provider. QoS requirements include:

Availability The availability of service is crucial for a company, whose interest it is to maximize availability of its data on a cloud. Cloud Service Providers promote cloud services by specifying uptime/downtime (99.5% uptime), offering guarantees and an uptime history. Do Cloud Service Providers offer Support Services: availability 24/7, live chat, online/self-serve resources, phone, webinars can be a decisive factor for customers in choosing the right Cloud Service Provider for their needs.

Security Data security is a major issue with Cloud Services. If customers typically avoid using Cloud Services it is because of data security reasons. As Schubert put it: "There is general distrust regarding outsourcing of data [1]." If a company's confidential data is transferred to a Cloud, data privacy and security are key areas to look

at. Trust in a Cloud Service Provider is enhanced, if data security is on their agenda and communicated. Phishing, data loss, password weakness [12], disaster recovery, firewalls, backup snapshots, privacy guarantee are all security related instruments implemented by Cloud Service Providers to ensure Cloud security. Cloud security is also tied the concept of privacy. It manifests itself for example in the cloud server's choice of datacenter location. For instance, German cloud users welcomed Amazon's expansion of Cloud Service datacenters to German territory, because they do not trust the United State's national data privacy policies [32]. Same as in security, there are many legal and technical problems associated with privacy, but privacy regulations can help mitigate the problem [1].

Scalability (elasticity)

- **Vertical**: scale up computing power within a single instance by increasing RAM or processing capacity without rebooting
- **Horizontal**: scaling to many users as defined in multi-tenancy, but customers are sensitive to whether the information about co-tenants in the Cloud is used or not [12] and this relates to privacy issues
- **auto scaling**: create automatic server load conditions that when met will trigger new virtual server instances to help carry the load without hindering performance (if CPU utilization goes above a certain threshold for a certain amount of time, a new instance will be created)
- **load balancing**: heavy traffic (traffic management) is automatically detected and rerouted across multiple instances to achieve greater levels of fault tolerance

Agility This criteria relates to the adaptability of the information on the cloud using one set of services to other sets services or servers. The ease of transporting data from one server to another when it is needed is important as the cloud user will not have to cater for the problems behind the provisioning of services [1].

Integrity Is the Cloud Provider known, publicly reviewed, listed at the stock exchange, sells promising and wanted shares, provides an amalgam of different services, states corporate profits in major newspapers, is listed on benchmarke websites? Provider reputation is an intangible criteria Cloud consumers use to assess the desirability of Cloud Service Providers. It is based on trust and reliability of the services provided [12].

The next two criteria, server os types and datacenter locations, are relevant if the following three criteria are fixed:

- Service Model: IaaS
- Deployment Model: Public
- Cloud Service: Computing

Regardless of what pricing models Cloud Service Providers use, the length of the contract period, customer segmentation or the quality of service the following criteria determine differences in prices in IaaS Computing.

4.3.1.7 Specifications

1. **Server OS Types** Linux (CentOS, Debian, Ubuntu, Red Hat Enterprise), SUSE Linux Enterprise, Windows Server Standard Edition, Windows Server 2008-2012, SQL Server Standard, SQL Server Web

2. **Datacenter Locations** Cloud customers can set up an account for all locations or choose to have more than one account for all kinds of locations. Depending on the Cloud Provider, datacenters are located worldwide or restricted to a region:

    - **continents**: Europe, North America, Asia, Australia, South America, Antarctica, Africa
    - **regions**: Western Europe, East Asia ...
    - **cities**: Frankfurt, Dublin, ...

4.3.1.8 Features

1. Number of instances or virtual machines (VM)
2. Instance name
3. Number of virtual CPU cores (vCPU)
4. Memory Power (GB RAM)
5. Storage (GB RAM)
6. System disk
7. Hard disk
8. Data disk
9. Bandwidth

4.4 Solutions

Based on the approach outlined in the previous section, different Cloud Service solutions will be explored in this section. Which Cloud Service Provider offers what cloud solutions? Three major and two minor Cloud Service Providers will be analyzed in more detail.
The Cloudscape on the internet grows day by day, since an increasing number of businesses are aware of the profitability of cloud solutions. Cloud Service Providers can ensue out of multinational big player companies, but there are also local cloud service providers. There are cloud service providers, which specialize in a cloud segment or outgrow other service providers to the extent that they become one of the major cloud providers on the market. To keep Cloud Service Providers comparable, only those which provide IaaS cloud services with datacenters located in Europe were investigated. The reason for this is that IaaS is the most comparable service level among all cloud services and the Cloudscape of Switzerland closest to year 2016 has the highest UZH-centric relevance in our time-space frame. Since big-player Cloud Service Providers operate in internationally relevant locations, Swiss locations most often do not host their datacenters. However, Western Europe has been the location of choice for building Amazon, Google or Microsoft cloud-related datacenters. This is why the location "Switzerland" was abstracted to the continent level, the European continent.

4.4.1 Cloud Service Providers

According to the Cloud Vendor Benchmark Switzerland IaaS 2015 by Experton Group, Amazon Web Services is the market leader and the most competitive Cloud Service Provider, which offers the most attractive cloud products [30]. It is by far and large the most well-known cloud service and also the most voluminous in terms of customer usage. Next, comparable leaders are Microsoft's brainchild cloud Microsoft Azure, Swisscom, T-Systems (relatively new on the market), Canopy, Google, IBM, BT, HP and Host Europe. The less competitive Cloud Service Providers challenging the market are Green.ch, ProfitBricks, VMware, NTT Comm., Interoute, MiroNet and CenturyLink. To name a few in low portfolio attractiveness and less than satisfying competitiveness, there is ORACLE and OVH.

4.4.2 Major Cloud Providers

Major Cloud Providers from a Swiss perspective are: AWS, Microsoft, Swisscom, T-Systems, Canopy, Google, IBM, HP, BT and Host Europe. For the sake of this paper, three major Cloud Providers will be compared, namely AWS, Google and Microsoft [30].

4.4.2.1 Amazon

Amazon has first started to leverage its infrastructure to deliver cloud computing services - the Amazon Web Services AWS - in 1997. Its advantage lies in the many cloud features and points of presence around the globe [11].
There are over 60 cloud services on AWS, which are grouped into 12 categories: compute, developer tools, Internet of Things, storage and content delivery, management tools, mobile services, database, security and identity, application services, networking, analytics, enterprise applications.
The customer can choose between four types of instances: on-demand, spot-instances, reserved instances and dedicated hosts. On-demand instances are paid on an hourly basis for only used instances. Capacity needs can be flexibly adjusted.

4.4.2.2 Google

Google has extended its services to cloud computing, the so-called Google Cloud Platform GCP with a focus on storage and network performance [11]. The Google Cloud Platform was launched in 2011 and serves on a worldwide basis. Google's headquarters are in California, USA. Prices are subject to automatic discounts with increased usage with no prepaid lock-in and per-minute billing. The Google Cloud Platform provides 16 different cloud service products: the App Engine, BigQuery, Bigtable, Compute Engine, Container Engine, Cloud Dataflow, Cloud Dataproc, Cloud Datastore, Cloud Pub/Sub, Cloud DNS, Cloud SQL, Cloud Storage, Prediction API, Stackdriver, Translate API, Vision API.

4.4.2.3 Microsoft

Microsoft is most famous for providing computer operation systems with an almost monopolistic position in its core business. Microsoft's headquarters are located in Washington, USA. It hosts the cloud platform Microsoft Azure, which was launched in 2010, with services ranging from analytics to computing, database, mobile, networking, storage and the web. Microsoft Azure incorporates IaaS, PaaS and SaaS with Linux and Microsoft Windows as Operating systems.

4.4.3 Minor Cloud Providers

Minor Cloud Providers on the Swiss market 2015 are Green.ch, ProfitBricks, VMware, Interoute, MiroNet, NTT Comm., CenturyLink, Oracle and OVH. Since green.ch does not offer an on-demand pricing scheme. ProfitBricks's server are not located in Europe. Oracle and OVH are not competitive and attractive enough on the IaaS market. Therefore, we chose MiroNet and VMware as minor Cloud Service Providers, because they have their datacenters in Switzerland or Germany (Europe) and offer an on-demand pay-as-you-go pricing scheme.

4.4.3.1 MiroNet

MiroNet was found in 2005. It is a Swiss IT company with headquarters located in Basel. They offer cloud-computing, IT and hosting services. MiroCloud offers elastic computing resources with guarantees of quality of service. Computing resources include CPU, memory storage, data storage and networks. A virtual data center disposes of the computing resources. Depending on the price model - pay-as-you-go or allocated - different pools can be created. In each of the pools a varying number of scalable virtual machines can be set by the customer.

4.4.3.2 VMware

VMware provides cloud infrastrucutre through the vCloud Air as part of the enterprise group Dell Technologies. Their headquarters are in Palo Alto, California. It was found in 1998. VMware's subsidiaries are located worldwide. They are mostly present in Europe. Their cloud architecture implements the concept of "one cloud, any application, any device". Their virtual infrastructure supports automation of computing, storage, networking and security services.

4.5 Evaluation and Discussion

In this section, pricing models are compared theoretically, followed by a realistic comparison of prices of three major and two minor Cloud Service Providers. Based on the criteria that were discussed above, the prices were pulled from the official online price list of the respective Cloud Service Providers. The data was gathered and evaluated by sorting the data according to ranks.

4.5.1 Comparison of Pricing Mechanisms

The difference between the fixed and variable pricing mechanism will be illustrated in illustrative examples below. As a preliminary give-away, the more sophisticated a model gets, the more input variables there are (price, hours, instances, vCPU, hard disk, bandwidth etc.) and the more input variables need to be addressed, the more complicated it is to calculate the costs. The input variables, for simplicity's sake were limited to two variables: prices and hours spent on a Cloud, everything else set to equal. To get a better approximation to the true value, this simplistic model is thought to have an added error term, i.e. an uncertainty variable, which would be a multiplier of some sort or a stronger overestimate of the average usage to make up for the the omitted variables.

1. Fixed price per month Green.ch offers customers a Virtual Cloud Server 1024 MB RAM, 50 GB SSD, 10 Mbit/s bandwidth for unlimited usage at a flat rate of CHF 49.90 per month.

Service	Provider	Product	Payment	Price(CHF)
IaaS	Green	Virtual Cloud Servers	Monthly	49.90

2. Variable price per month Microsoft Azure offers customers Virtual Machines such as D1 instance, 1 vCPU, 3.5 GB RAM, 50 GB hard disk with pay-as-you-go at an hourly rate of USD 0.01 (744 hours maximum a month). Customers are charged only for what they use. If they use the virtual machine for 150 hours in the first month, the price will be CHF 0.134*150 = CHF 20.10 and so forth.

Service	Provider	Product	Payment	Price(CHF)
IaaS	Azure	Virtual Machines	Monthly	See table below

Month	Hours used	Hourly rate(CHF)	Payment	Price(CHF)
1. Month	150	0.134	Monthly	20.10
2. Month	200	0.134	Monthly	26.80
3. Month	744	0.134	Monthly	99.70
Total	**1020**	**0.134**	**Season**	**146.60**

From a financial standpoint, a comparison of different payments in time is usually done by discounting the sum of payments, also known as discounted cash flow (DCF). The discount factor is usually derived by opportunity cost, often a risk-free rate, which is currently close to zero or negative. Another approach would be to use the company's ROE, since money kept in the company could be used elsewhere in the company. One could also look at Moore's law that states a 25% price reduction annually, which is quite high and only true for equal computing power. Since, the demand has steadily increased and probably will in the future cash flow will probably not be affected by it. However, the DCF model is used to compare investments and the resulting cash flows. The goal is to be able to compare them in the same period, the present value. If there are no setup cost involved and prices are only paid monthly, we can directly compare the monthly payments, as the difference today will only get bigger over time. An hourly rate will scale linearly with each additional hour, while a flat rate stays flat. To determine the break-even point the flat rate has to be divided by the hourly rate and we will receive their intersection. For example if an hourly rate is CHF 0.5 and a flat rate CHF 49.90 it is better to use an hourly basis up until the 99.8 hour.
For reserved instances, such as in the AWS we can make a similar comparison. Many companies would rather choose predictable payments than risky, uncertain payments, even if such uncertain payments were lower priced.
If the fixed flat rate is charged at CHF 49.90 as in Green.ch and AWS is used as cloud service provider for instance m3.medium, 1 vCPU, 3.75 GB RAM charged at an hourly rate of CHF 0.144, the following model will hold: The break-even point, at which the cloud user is indifferent is at 372 hours. The cloud user fares better with a reserved instance pricing mechanism, beyond this 372 hours, a fixed flat rate pricing mechanism makes more sense. Although the hourly pricing schemes seems cheapest compared to the monthly flat rate charging, the cloud provider will still pay for whole hours, even if less is used. The logic monthly to hourly could be extended to the minute-level: if a cloud provider were to charge cloud users on a minute basis the incidental charges would make a logarithmic difference. If a cloud user were to use 5 minutes on a virtual machine, the price charged for AWS would be the one-hour (hourly) rate. However, if a cloud provider were to charge on a minute level, 55 minutes of overcharge could have been avoided. For spot instances (AWS) the pricing scheme is built on an auction basis: consumers place a maximum bid on a virtual machine within a short period of time. If the bid is above the current spot price the consumer gets access to the resources. They will not be

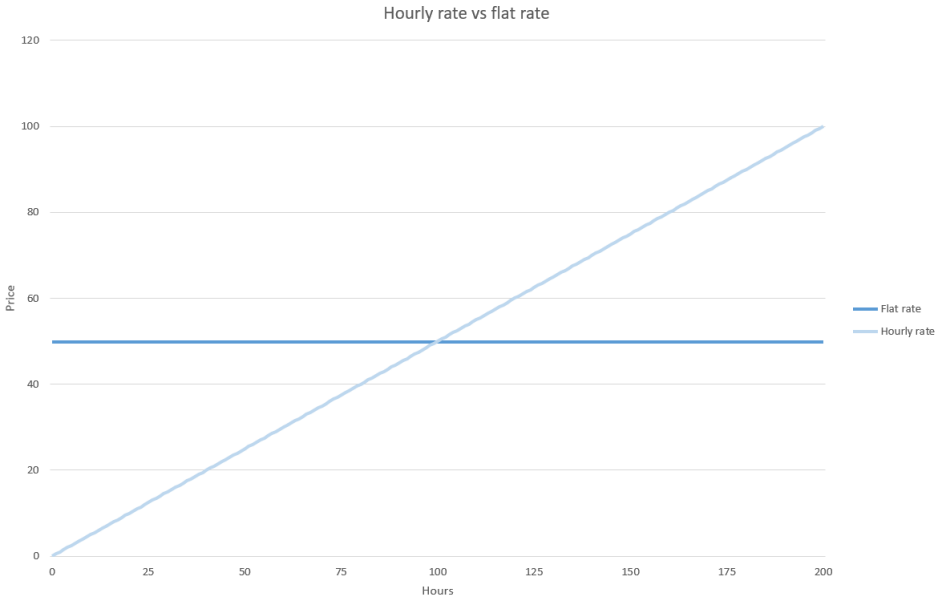


Figure 4.2: Flat rate vs Hourly rate

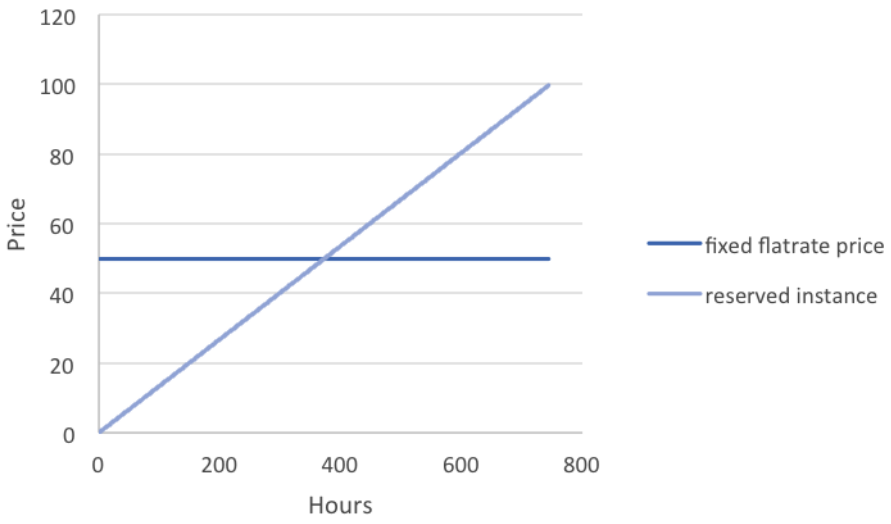


Figure 4.3: Green vs AWS

charged for whole hours and are therefore cheaper than hourly payments. The downside of this scheme is that consumers can be kicked off by AWS at any time. Which Cloud Providers are best? There are problems with the cloud kept in secret such as performance issues, security issues, bottlenecks, high rate of disk failure. Hence, we recommend to check internet cloud brokers such as Cloudorado for availability, quality, performance and customer service.

4.5.2 Comparison of Prices

Where prices were not shown in Swiss francs, the foreign currencies were transformed into Swiss francs. This was done by using two foreign exchange rates, the USD-CHF foreign exchange rate at 0.903 and the EUR-CHF foreign exchange rate at 1.078.

4.5.2.1 First Configuration

- vCPU: 1

- Memory/Storage GB RAM: 3.75
- Operating System: Windows

Ranking:

Rank	Provider	Price
1	GCP	41.62 CHF/month
2	VMware	89.13 CHF/month
3	AWS	93.88 CHF/month
4	Microsoft Azure	106.54 CHF/month
5	MiroNet	408.75 CHF/month

Where prices were tied to varying GB RAMs unequal to 3.75 GB RAM such as 3.5 GB RAM, a linear interpolation was applied to determine the price of 3.75 GB RAM. The closest GB RAM to 3.75 served as a basis for all Cloud Service Provider offerings.

4.5.2.2 Second Configuration

- vCPU: 4
- Memory/Storage GB RAM: 16
- Operating System: CentOS or Ubuntu Linux

Ranking:

Rank	Provider	Price
1	VMware	82.00 CHF/month
2	GCP	101.52 CHF/month
3	AWS	188.40 CHF/month
4	Microsoft Azure	182.76 CHF/month
5	MiroNet	195.50 CHF/month

GB RAM ranged from 14 to 16 for 4 vCPUs. To increase the prize for 14 and 15 GB RAM a simple proportional linear calculation based on the rule of three was performed.
Criteria that apply to all five IaaS Cloud Service Providers at the same time was difficult to find. For example, if the operating system changed from Windows to Linux for only one single vCPU, Microsoft Azure would have been excluded because Linux is only available with a minimum of 4 vCPUs placed in an instance.
Hence, it can be deduced that provided that all things are equal (ceteris paribus), among the major Cloud Providers, GCP is cheaper than AWS. AWS, in turn, is cheaper than Microsoft Azure in both scenarios.

1. GCP
2. AWS
3. Microsoft Azure

The prices of the minor Cloud Providers VMware and MiroNet are quite different. MiroNet in both cases charges the highest price - in the first case it was substantially higher. MiroNet is the only Cloud Provider with a corporate residence. The Swissness could explain the high prices that were charged.
VMware ranks highest in both cases. It holds either first or second position. Hence, the size of the Cloud Service Provider can be said not to have an influence on the prices. But further comparisons must be made to have representative conclusions of the findings.

1. VMware
2. MiroNet

Based on this limited study of prices, the white paper published by a Google sponsored research group was right in claiming that GCP prices were lower than AWS prices. This leaves out the possibility of comparing reserved instances, because Google does not apply as many pricing models as AWS does.
The major Cloud Service Providers can be compared against each other as well. The criteria applied for the major Cloud Service Providers do not exist for minor Cloud Service Providers. Hence, the evaluation is limited to the major Cloud Service Providers Google, Amazon and Microsoft. The number of vCPUs and memory/storage GB RAM remains equal. Changes were applied to the operating system. First SUSE was compared, followed by SQL Server Standard, SQL Server Web and the Red Hat Enterprise Linux Operating System.

4.5.2.3 Third Configuration

- vCPU: 4
- Memory/Storage GB RAM: 16
- Operating System: SUSE

Ranking:

Rank	Provider	Price
1	GCP	108.29 CHF/month
2	AWS	254.51 CHF/month
3	Microsoft Azure	362.45 CHF/month

4.5.2.4 Forth Configuration

- vCPU: 4
- Memory/Storage GB RAM: 16
- Operating System: SQL Server Standard

Ranking:

Rank	Provider	Price
1	AWS	254.51 CHF/month
2	GCP	683.51 CHF/month
3	Microsoft Azure	719.52 CHF/month

4.5.2.5 Fifth Configuration

- vCPU: 4
- Memory/Storage GB RAM: 16
- Operating System: SQL Server Web

Ranking:

Rank	Provider	Price
1	GCP	251.75 CHF/month
2	AWS	321.94 CHF/month
3	Microsoft Azure	436.93 CHF/month

4.5.2.6 Sixth Configuration

- vCPU: 4
- Memory/Storage GB RAM: 16
- Operating System: Red Hat Enterprise Linux

Ranking:

Rank	Provider	Price
1	GCP	199.71 CHF/month
2	AWS	228.06 CHF/month
3	Microsoft Azure	254.94 CHF/month

In all configurations but one, GCP ranks highest with the lowest price, followed by AWS. Microsoft Azure charges the highest price throughout each scenario.
Hence, GCP can be said to have the overall lowest price across all six different configurations made. AWS ranks second and Microsoft Azure ranks third.

4.5.2.7 Explanation

All chosen Cloud Service Providers are based on the Swiss IaaS Cloudscape benchmark 2015 [30]. For effective comparison between Cloud Service Providers, the same price-determining criteria must apply to each one of them. Assuming that certain criteria have a certain value, maximum equalization among different criteria was the goal.

- **Service Model**: All Cloud Service Providers provide IaaS services. IaaS is most comparable among all Service models.
- **Deployment Model**: All Cloud Service Providers host public clouds.
- **Pricing Model**: All Cloud Service Providers have on-demand (pay-as-you-go), hence an hourly price rate and a monthly flat-rate price. The comparison will be based on monthly prices charged (converted to, if necessary) in Swiss francs. The monthly price (31 days a month at full usage) will not be paid less hours were paid. In order to make offerings comparable, it is assumed that the computing resources ran a maximum of 24 hours a day, 31 days a month at full capacity. Hence, the monthly prices are comparable to each other.

As a side note, AWS and VMware offer another pricing model: a subscription based payment method, where the subscription terms are divided into the subscription period (starting from 1 months and up) and payment method (monthly or prepaid). Prices differ across pricing models. Comparison within a pricing model makes sense. Considering the scope of this paper, only one pricing model will be analyzed. Servers, which are shared among multiple tenants, are the object of the evaluation. AWS and VMware have dedicated servers, too. But reserved or dedicated cloud server instances fall out of the scope of the analysis because they belong to another pre-paid subscription based pricing model with corresponding discounts.

- **Contract Period**: one month of full, flat-rate price payed for equal Cloud services. If the contract period had begun in a trial period with no prices charged, it would not have been comparable.
- **Customer Segmentation**: irrelevant to the Cloud Service chosen
- **Cloud Services**: Cloud Computing virtual machines infrastructure with no additional services included

- **Quality of Service**: all services set to zero to ensure no additional impact on price variability due to better quality of service
- **Specifications/features**: different configurations were made based on Cloud Service Provider comparability, however with regards to location all Cloud Service Providers have datacenters with servers located in Europe. The number of instances was limited to one for reasons of simplicity.

4.6 Summary and Conclusion

In this section, a brief summary of the findings is provided and conclusive remarks are made.

4.6.1 Summary

First, Cloud Computing was defined in relation to the related components of Cloud Computing and the problem statement of which Cloud Service Provider is most affordable with regards to emerging pricing models was presented. Second, the literature on Cloud Service Providers was reviewed. Third, the most important approaches to Cloud Services were discussed and criteria advanced for the assessment of different Cloud Services with respect to the prices charged. Forth, a set of major and minor Cloud Service Providers were set forth and brought into context. Finally, the evaluation of a price comparison of configured Cloud Services from the determined set of Cloud Service Providers was laid out and the results were expounded upon.

4.6.2 Conclusion

First, the results of this paper show that the Google Cloud Platform is catching up by providing cheaper Cloud Services than Amazon Web Services. Hence, the monopolistic market position and popularity of AWS should not hold back emerging Cloud Service Providers such as T-systems from entering the cloud market.
Second, explanations for MiroNet's high-end Cloud Service offerings were given. The high prices charged by MiroNet could be due to the Swissness or the small size of the company. Finally, the results imply that although GCP is second to AWS and offers cheaper prices, the same yardstick cannot be applied to minor Cloud Service Providers. VMware operates internationally, but MiroNet is mostly contained to the Swiss market. Minor Cloud Service Providers are both expensive and do not offer as many additional and different services and models. Hence, the competitiveness is most difficult if Cloud Service Providers are still in the burgeoning phases. However, once the pool of big players is reached, potential for competitiveness among Cloud Service Providers can be tapped into. This is also due to the relative novelty of the market sector.

4.6.3 Threats to Validity

The results of this study are exposed to risks of validity of an external and internal nature. The external validity deals with threats and potential pitfalls, which were identified as emanating from the environment. The internal validity addresses the issues with internal factors that arose from producing the results for this paper.

4.6.3.1 External Validity

Since only a limited number of configurations were looked at, the generalization of this study might be limited. As long as Cloud Service Providers choose to set prices idiosyncratically, the harder it gets to draw useful comparisons among Cloud Service Providers. Nevertheless, attempts were made to mitigate the risk of invalid results. IaaS was chosen for this study because it is the most comparable level of all service models. Another threat to generalize is the selection of Cloud Service Providers. Most of the Cloud Service Providers operate on an international scale, even if they are smaller in size (VMware) compared to the big players (Google). This was remedied by sticking to the Swiss Cloudscape and controlling for different other criteria, which resulted in a fewer number of choices among Cloud Service Providers.

4.6.3.2 Internal Validity

Cloud prices were pulled from the official homepages of the Cloud Service Providers within the last few weeks. It is not known to what extent Cloud Service Providers update the prices published on their website and whether this time interval is short enough to invalidate this study's findings already right upon its completion. If Amazon were to know of GCPs cheaper offerings, it might react any time to offer even cheaper cloud services. However, to this moment it can be safely said that this was not the case. The results capture only a snapshot of the current cloud market reality.

4.6.4 Outlook

Cloud Computing is becoming increasingly relevant for cloud users such as IT companies and Cloud Service Providers alike. It is imperative to further investigate the pros and cons of pricing models in Cloud Computing. But it can be said for sure that there will never be one best pricing model. This is due to the differences in requirements - ad hoc or continuous - of cloud users and corresponding cloud services. The more complex the requirements of users and the bigger the capacity to pay, the more customized and individualistic the cloud services get. The pricing is often an agreement (SLA) resulting from price negotiations between cloud consumers and Cloud Service Providers within the scope of legal security and privacy regulations. Likewise, it is important to gain an understanding of the reasons why different Cloud Service Providers charge different prices for the same services. This will help cloud users challenge Cloud Service Provider's pricing policy and relative market power reflected in cloud service SLAs and help to derive a useful tool of comparison for making sense and navigating through a myriad of different cloud services found on the Cloudscape. This paper has made contributions to each of these areas. Further research needs to be made with regards to different configurations of on-demand and subscription-based pricing benchmarks and other Cloud Service Providers than the ones used in this study.

Acknowledgement

We thank green.ch and MiroNet employees for their time taken to give us advice and information on their Cloud Services and pricing models.

Bibliography

[1] L. Schubert, K. Jeffery: *Advances in Clouds*, European Commission, Publications Office of the European Union, Luxembourg, 2012. http://cordis.europa.eu/fp7/ict/ssai/docs/future-cc-2may-finalreport-experts.pdf

[2] Philipp Leitner, Jürgen Cito: *Patterns in the Chaos - A Study of Performance Variation and Predictability in Public IaaS Clouds*, ACM Trans. Internet Technol. 16, 3, Article 15, Vol. 16, Issue 3, Article No. 15, New York, USA, August 2016. http://dl.acm.org/citation.cfm?id=2885497

[3] Wikipedia: Cloud Computing; https://de.wikipedia.org/wiki/Cloud_Computing, Month, 2016.

[4] Amazon Web Services; https://aws.amazon.com/de/, Month, 2016.

[5] Google Cloud Platform; https://cloud.google.com/, Month, 2016.

[6] Microsoft Azure; https://azure.microsoft.com/de-de/, Month, 2016.

[7] Cloud Provider; https://cloudprovider.net/, Month, 2016.

[8] 1&1; https://hosting.1und1.de/dynamic-cloud-server, Month, 2016.

[9] May Al-Roomi, Shaikha Al-Ebrahim, Sabika Buqrais, Imtiaz Ahmad: *Cloud Computing Pricing Models: A Survey*, International Journal of Grid and Distributed Computing, Vol.6, No.5 (2013), pp.93-106.

[10] Se-Hak Chun, Byong-Sam Choi: *Service models and pricing schemes for cloud computing*, Springer Science+Business Media New York, 2013.

[11] Aviv Kaufmann, Kerry Dolan, ESG Lab White Paper: *Price Comparison: Google Cloud Platform vs. Amazon Web Services*, The Enterprise Strategy Group, Inc., June 2015. https://cloud.google.com/files/esg-whitepaper.pdf

[12] Artan Mazrekaj, Isak Shabani, Besmir Sejdiu: *Pricing Schemes in Cloud Computing: An Overview*, (IJACSA) International Journal of Advanced Commputer Science and Applications, Vol. 7, No. 2, 2016. http://thesai.org/Downloads/Volume7No2/Paper_11-Pricing_Schemes_in_Cloud_Computing_An_Overview.pdf

[13] Hong Xu, Baochun Li: *Maximizing Revenue with Dynamic Cloud Pricing: The Infinite Horizon Case*, University of Toronto, 2012. http://iqua.ece.toronto.edu/papers/hxu-icc12.pdf

[14] Varun Kamra, Kapil Sonawane, Pankaja Alappanavar: *Cloud Computing and Its Pricing Schemes, International Journal on Computer Science and Engineering*, Sinhgad Academy of Engineering, Pune, India, Vol. 4, No. 04, April 2012. http://www.enggjournals.com/ijcse/doc/IJCSE12-04-04-127.pdf

[15] Sushil Bhardwaj, Leena Jain, Sandeep Jain: *Cloud computing: A study of infrastructure as a service(IaaS)*, (IJACSA) International Journal of Engineering and Information Technology, 2010.

[16] Parnia Samimi, Ahmed Patel: *Review of pricing models for grid & cloud computing*, IEEE, 20-23 March 2011.

[17] Peter Mell, Timothy Grance: *The NIST Definition of Cloud Computing*, NIST Special Publication 800-145, September 2011.

[18] Von Neumann architecture; `https://en.wikipedia.org/wiki/Von_Neumann_architecture`, October, 2016.

[19] Helmut Herold, Bruno Lurz, Jürgen Wohlrab: *Grundlagen der Informatik*, Pearson Deutschland GmbH, 2012.

[20] Von Neumann architecture picture; `http://s1141.photobucket.com/user/Muhammad_Soban/media/Von\%20Neumann\%20architecture/vonNeumannComputerModel.jpg.html`, October, 2016.

[21] Peter Rechenberg, Gustav Pomberger: *Informatik Handbuch*, Carl Hanser Verlag München Wien, 2006.

[22] Virtualization in Cloud Computing; `http://eisct.in/Adminfiles/(201503)VIRTUALIZATION\%20IN\%20CLOUD\%20COMPUTING.pdf`, November, 2016.

[23] Multicore Virtualization over a Multikernel; `http://e-collection.library.ethz.ch/eserv/eth:7128/eth-7128-01.pdf`, November, 2016.

[24] Pankesh Patel, Ajith H. Ranabahu, Amit P. Sheth: *Service Level Agreement in Cloud Computing*, The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar, Kno.e.sis Publications, 2009. `http://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1077&context=knoesis`

[25] Christian Gronross: *Service Quality: The Six Criteria Of Good Perceived Service Quality*, Review of Business, Winter, 1988.

[26] ISO/IEC 25010; `https://www.iso.org/obp/ui/#iso:std:35733:en`, November, 2016.

[27] On-premises software; `https://en.wikipedia.org/wiki/On-premises_software`, November, 2016.

[28] Foster, I. Zhao, Y., Raicu, I. and Lu, S.: *Cloud Computing and Grid Computing 260-Degree Compared*, 2008 Grid Computing Environment Workshop (GCE '08), 2008.

[29] Pragya Gupta, Sudha Gupta: *Mobile Cloud Computing: The Future of Cloud*, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2012.

[30] Heiko Henkes, Frank Heuer, Oliver Giering, Oliver Schonschek, Wolfgang Heinhaus und Arnold Vogt: *Cloud Vendor Benchmark 2015 - Cloud Computing Anbieter im Vergleich Schweiz*, Experton Group AG, 2015. `https://www.t-systems.com/blob/55422/07db03ee16c01e81996745e024017355/2015-06-30-pdf-experton-benchmark-data.pdf`

[31] Michael Essery: *Latest 451 Research Cloud Price Index Reveals a 74% Saving by Using Multiple Cloud Providers*, 451Research, 2015. https://451research.com/images/Marketing/press_releases/11.19.15_CPI_Q4_FINAL.pdf

[32] Was AWS, Microsoft und Co. in Deutschland vorhaben; http://www.computerwoche.de/a/was-aws-microsoft-und-co-in-deutschland-vorhaben,3070953, November, 2016.

Chapter 5

Different Impacts of the New Swiss Law on Monitoring of Postal and Telecommunication Traffic

Olga Klimashevska, Nicola Staub, Jonas Wagner

In March 2016, the Swiss parliament adopted a new version of the federal law concerning the monitoring of postal and telecommunication traffic (free translation of "Bundesgesetz betreffend der Überwachung des Post- und Fernmeldeverkehrs"), BÜPF for short. New regulations always come together with critical voices. Taking three different perspectives, this report analyzes what impact this new and extended law has on the Swiss citizens. Fast evolving technologies make it harder for intelligence and prosecution agencies to monitor criminal perpetrators, calling for more sophisticated software products from a technical perspective. Increased surveillance also affects the providers of postal or telecommunication services, as they are obliged to invest in means for surveillance as a consequence. The report investigates the implications of these costs, different providers faced up to now, and what expenses can be expected when the new regulation becomes effective in January, 2018. The report highlights the main areas costs are increasing in the future, whereby small and medium sized companies suffer the most due to high investment costs. The influence on a person's behaviour, being at the mercy of surveillance, is critically analyzed from a social viewpoint. As there might be violations of privacy or data protection when monitoring people at large scale, the BÜPF is legally scrutinized with regard to security and social acceptance.

Contents

5.1 Introduction

Imagine the following situation: Dr. Bob's neighbour Alice is a successful cyclist, who has already won several national races and is also competing at an international level. She is a friendly and helpful person, Dr. Bob gets along very well. Alice has been thankful for medical advices of Dr. Bob multiple times already. Yesterday evening she was calling him, inviting him for dinner. During this call, Dr. Bob was also wondering how Alice was doing at the race she competed few days ago.
As the authorities are accusing Alice for taking advantage of illegal substances for performance enhancement, she is currently monitored by the authorities. All her postal and telecommunication traffic gets monitored and checked for collecting evidence. The telephone conversation Dr. Bob had with Alice also got intercepted, possibly making him and his medical practice a target of the ongoing doping investigation as well.
Is it legitimate to intercept phone calls from a suspect, even when hazarding the consequences of endangering the confidential medical secret? What about the right for privacy and data protection?
It is certainly hard to find a standard answer for these questions, as the circumstances are heavily intertwined. This report tries to answer these questions by looking at the BÜPF from different perspectives. After summarizing the evolution and scope of this law, implications from a technical perspective are highlighted. The second part consists of an analysis of costs implied by the BÜPF, with regard to its economic environment. In the subsequent part, the report looks at the BÜPF from a social perspective, including psychological or legal ramifications. Finally, correlations between the different perspectives are shown and compared to regulations of other countries.

5.2 What is the BÜPF?

The "Bundesgesetz betreffend der Überwachung des Post- und Fernmeldeverkehrs" (short *BÜPF*) is a collection of laws, on which the Federal Assembly and the Federal Council of Switzerland has decided on in March 2016. In the upcoming version of the BÜPF, there will be major changes in comparison to the previous versions. The reason why the Federal Assembly has changed the law is mainly because of the vast development of communication technologies [44]. This means, that technological progress in electronic communication has made it increasingly difficult for the federal legal enforcement institutions to monitor people. Therefore, according to the Federal Council, those institutions need to be supported by extending their legal abilities and by creating a centralized service for monitoring postal and telecommunication traffic.

5.2.1 Intended and Official Purpose of the new Version of the BÜPF (March, 2016)

The official purpose of the BÜPF is to regulate the monitoring of postal and telecommunication traffic, in order to find, identify or investigate on persons of interest. The law regulates the legal boundaries, which allow federal legal enforcement institutions the use of certain instruments to monitor people inside Swiss borders. There are four main targets areas of monitoring regulated by the BÜPF [37, 38]. First, monitoring a person in case of a criminal procedure. Second, a person can be monitored in order to enforce judicial assistance. Third, finding a missing person, and fourth, tracking of already condemned people, to enforce sentences, causing deprivation of liberty. All surveillance actions, which are taken by prosecution authorities, need to conform with one of the previous targeted ar-

eas. Therefore, the BÜPF is primarily an instrument to support the criminal prosecution authorities.
Although the official purpose of the BÜPF is clearly defined, there are reasons why the Federal Council, the State Council and the National Council decided to revise the current version of the BÜPF. One argument was, that technological progress of the digital industry has created loopholes for criminals, which shield them from surveillance [43]. Therefore, the new BÜPF should facilitate the situation of criminal prosecution authorities by allowing them to use state-of-the-art tools and processes, which are able to overcome the digital hiding places of criminals under surveillance. Another reason why the Federal Assembly and the Federal Council wanted to revise the BÜPF, is to simplify the workflow for the Central Service BÜPF (CSB) (free translation of "Dienst ÜPF") [44, 45]. This would allow a more effective and less time consuming setup for surveillance. In the past, some criminal prosecution authorities already used monitoring tools, which had no legal base in the Swiss law. The new bill should legalize those methods. According to the Federal Council, a further reason supporting a revision of the BÜPF, was the intention to give the authorities more data in order to fully enable them to monitor and catch criminal perpetrators under surveillance.

5.2.2 History/Development of the BÜPF

The Swiss BÜPF is based on the federal laws concerning the monitoring of telephone traffic and telegraph traffic (free translation of "Schweizer Telegrafen und Telefonverkehrsgesetz"), which were passed by the Federal Assembly in October 1922 [47]. Before that time, there was no federal law, which regulated surveillance in Switzerland. The state law defined, under which circumstances a person can be monitored. Later, in 1942, the laws were complemented by adding a paragraph, which released the former national Post-Telefon-Telegram Company (PTT) from the secrecy of Post in case of monitoring. In 1991, the laws concerning the monitoring of telephone and telegraph traffic were replaced by a collection of laws called Swiss telecommunication law (free translation of "Schweizerisches Fernmeldegesetz"). The main difference between the Swiss communication law and its predecessor was, that the monitoring of telephone calls are regulated on a national base from this point on. Previously, the Swiss Post was only explicitly released from the secrecy of service (e.g. secrecy of mail).
In 1993, the Federal Assembly and the Federal Council decided to split up PTT into two privatized companies [26, 47]. Those companies are known as the Swiss Post, which took care of the postal traffic, and Swisscom, which took responsibility for the telephone and upcoming Internet traffic. The privatization led to difficult judicial situations, because each employee of the PTT underlied the obligation of secrecy of officialdom. This obligation would not be valid anymore in a private company. Therefore, in 1997, the Swiss telecommunication law was revised and prepared the foundation of the BÜPF. Parallel to the telecommunication law, the very first version of the BÜPF was introduced, thus, the Federal Assembly had to regulate not only the postal, telephone and telegraph traffic, but also newer kinds of communication like Internet traffic. In addition to the BÜPF, a service was established, which was responsible for all state requests for monitoring of either postal or telecommunication traffic.
Until the subsequent version of the BÜPF (passed in October 2000) [37, 47], the different cantons decided themselves, whether monitoring a person is justified or not. The version of the BÜPF in year 2000 introduced a catalogue, which defined all criminal actions and situations in which surveillance was justified [16]. These criminal actions and situations reached from rather small crimes (e.g. theft, doping, material damage) to really serious crimes (e.g. financial support of terrorism, pedophilia, genocide). Further, the telecom providers were forced by law, to store metadata (free translation for "Randdaten") of every

customer and their actions for at least six months. An upper limit did and still does not exist. In addition to that, all requests for surveillance had to have a judicial approval. In 2007, the Federal Council founded a central service (CSB), which is responsible for supervising and proper execution of all monitoring requests from the cantons. Later, in 2011, the competences of the CSB and the authorities of criminal prosecutions, which differ for every canton in Switzerland, were adapted. The mandatory judicial approvement, whenever a crime over the Internet took place, was removed. In 2012, the government forced all telecommunication providers by legal ordinance (but without legitimate base) to record full data streams of their customers.

5.2.3 Comparison of the BÜPF (July, 16, 2012) with the latest Revision (March, 18, 2016)

The BÜPF of 2012 roughly counts about 1'920 words [37]. This is less than half the size compared to the upcoming BÜPF (ca. 4'841 words) [38], which will be in act on January 1st, 2018 [27]. Since it is not beneficial to cover every single change of the two versions of the law, this section focuses only on the most important parts.

In the version of the BÜPF from 2012, the CSB only had compentencies in two areas: First, supervising surveillance requests, and second, the responsibility for proper execution of surveillance procedures. With the new version of the BÜPF, the compentencies of the CSB will be extended further. From 2018 on, the service will be responsible for the Federal Informatics System (free translation of "Informatiksystem zur Verarbeitung von Daten im Rahmen der Überwachung des Fernmeldeverkehrs") for processing data in the context of the BÜPF. According to the BÜPF 2018, the informatics system is specified as a system with a database, tools for filtering and processing data, and a data interface for the different legal enforcement institutions. In this report, this informatics system is referred to by the term *IT-Service BÜPF* (ISB) [38].

In addition to the new ISB, there is a change of the procedure about how the information is collected from providers of electronic communication. Up until January 2018, the CSB gave orders to providers once a request for monitoring was proven valid. Afterwards, the involved providers began to provide their data to the CSB. From 2018 on, providers (and a few other companies, which are defined by the Federal Council) have to be ready at any time, to deliver their collected data at a certain standard. This standard will be enforced by granting a "proof of preparedness" to all affected providers of electronic communication (and the additional companies defined by the Federal Council). The affected companies have to be financially liable for the entire process for getting such a proof. If a company is not ready to deliver data or is hiding data from the government, it can be sanctioned with a fine up to CHF 100'000 [37, 38].

If a request for monitoring causes actions which are not standardized according to the proof of preparedness, the CSB has to choose the "easiest way of monitoring the target". This means, that the CSB chooses a provider, which is likely to be the fastest one to deliver the required data. All other providers, which are involved in the monitoring process, have to deliver their data only to the provider chosen by the CSB. The latter also have to grant immediate access to their assets to the chosen provider [37, 38].

The Federal Council introduced an article in the new BÜPF, which invalidates all laws of data privacy protection in case of surveillance. This means, that the CSB and the law enforcement institutions have something like a "blank check for processing and collecting data" as long as they have a mandate to do so [37, 38].

From 2018 on, every person or company, which is responsible for an access point to the Internet, television or telephone network, not only has to tolerate monitoring, but also has to deliver all available metadata to the CSB or the responsible provider chosen by the

CSB. Additionally, those people or companies have to help the responsible authorities to monitor and collect data, by granting them access to their assets [37, 38].
Other important changes are related to different collections of laws, which are caused by the extension of the BÜPF. In the past, some criminal prosecution institutions used so called IMSI-Catchers, antenna scans (free translation of "Antennensuchlauf") and government supported malware called "GovWare", to monitor persons of interest [47] (c.f. Section 5.4). In the case of antenna scans or IMSI-Catchers, the federal court has subsequently legitimated their use, although there was no legal scope to do so [21]. With the new BÜPF, the infiltration of data processing devices with special software (free translation for "besondere Informatikprogramme") is legitimated. This allows the government to produce trojan horses and other malicious software to monitor persons of interest [37, 38]. As a side note, the duration of storage for metadata (data retention) remains unchanged at a minimum of six months [37, 38]. The Council of States prevented the Federal Assembly and the Federal Council from extending this storage period [45].

5.2.4 BÜPF and its Dependent Services

In order to explain the future interconnection of the different services, institutions and actors of the new version of BÜPF, the report introduces a fictive case study.
Giovanni is a leader of a gang, which is heavily involved in organized criminal actions like human-trafficking and illegal drug distribution. He has running subscriptions on several telecommunication and Internet providers, in order to disguise his communication channels from the government. He owns a notebook, several TVs and two mobile phones. Giovanni has a liaison with a woman called Francesca. It is assumed that Francesca does not know about Giovanni's criminal affairs. Further, the Swiss federal prosecutor Hans has been investigating Giovanni's gang for some time already. Since Hans has enough evidence of Giovanni's criminal actions, he now orders surveillance measures from the CSB, in order to chase down Giovanni and his gang. For the sake of simplicity, the case study will not cover the surveillance process of postal traffic, because there is no significant difference in monitoring this kind of traffic in the current and upcoming version of the BÜPF in 2018.

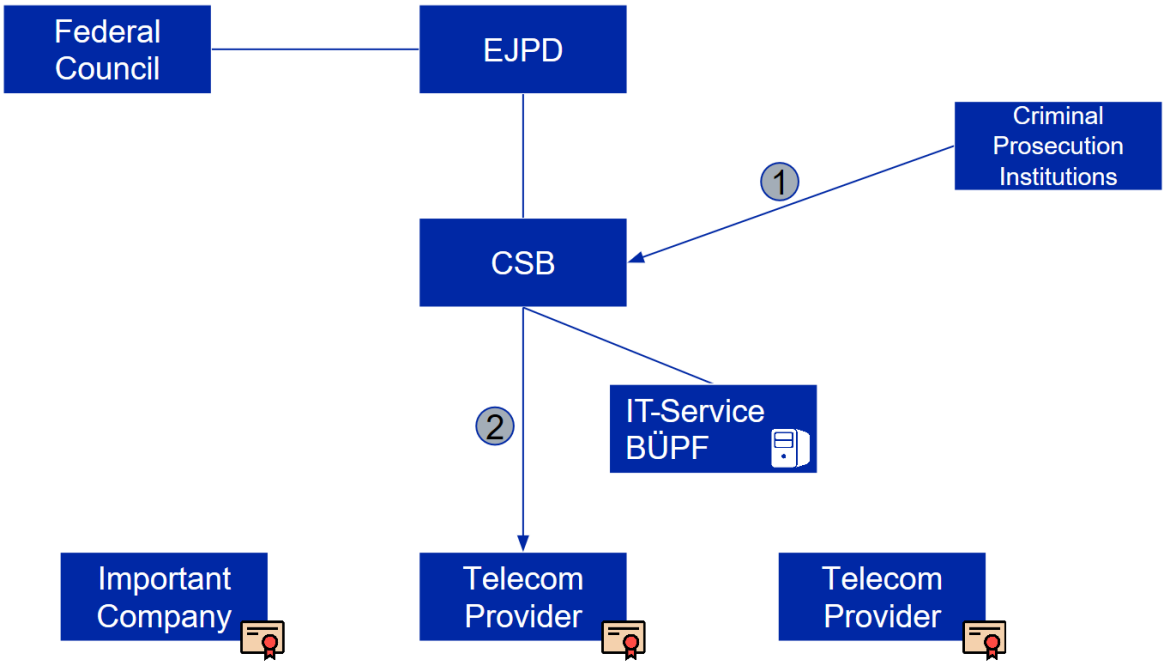


Figure 5.1: Process of surveillance part one (own Graphic).

The Figures 5.1 and 5.2 depict the involved parties in the upcoming version of the BÜPF and illustrate the steps of the case study which are going to happen. The Federal Department of Justice and Police (free translation of "Eidgenössisches Justiz- und Polizeidepartement", short EJPD) thereby acts as the link between the Federal Council and the CSB.
Step 1: Once the federal prosecutor Hans has the concession from a state judge or another authorized institution, he contacts the CSB in order to initiate the monitoring process [38]. In a second step, the CSB checks the permission of Hans and, if it is valid, starts the monitoring process by deciding, which of the affected telecommunication providers is the fastest one to deliver the required information. The reason why the CSB needs to decide for one single telecommunication provider is due to the fact, Giovanni uses several different providers. After the CSB has chosen a telecommunication provider, this provider contacts them, and orders the stored metadata of Giovanni and his environment, as well as all other available data.

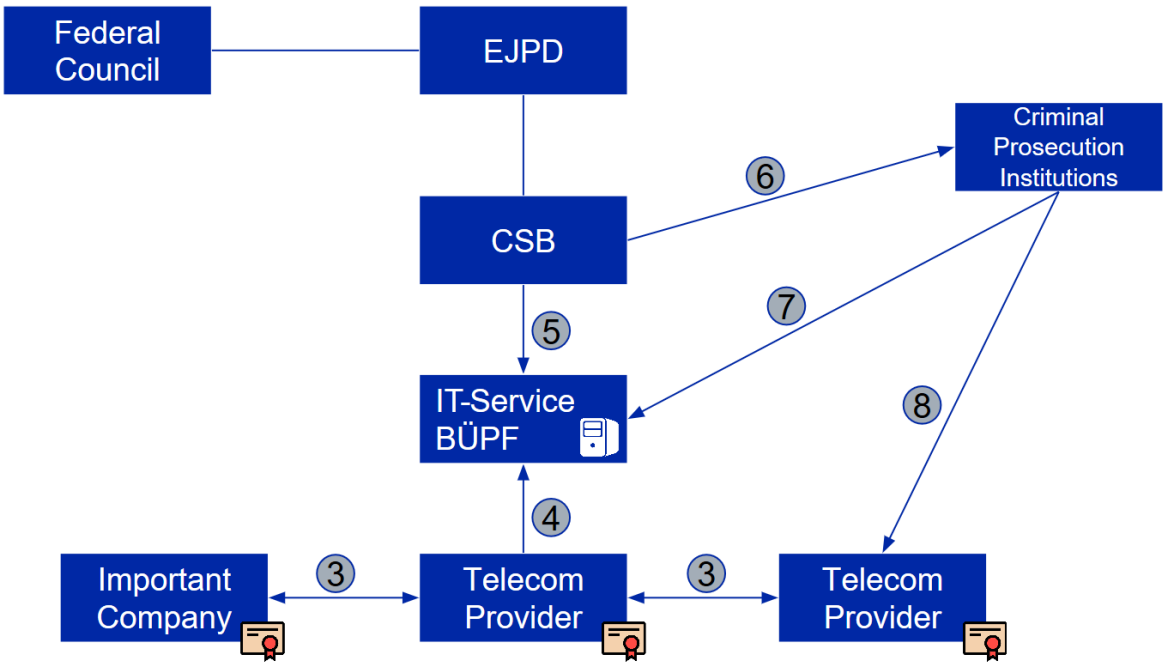


Figure 5.2: Process of surveillance part two.

Since all telecommunication providers and the companies chosen by the Federal Council have to be ready for surveillance at any time, the chosen telecommunication provider begins to deliver Giovanni's data to the ISB immediately. At the same time, this provider contacts all other involved providers and companies, to collect the available data about Giovanni and his environment from them. These actions are illustrated in step three and four of Figure 5.2. As an important side note, all costs, which are caused by the surveillance of Giovanni, are paid by the involved providers and companies themselves. Additionally, since Francesca had contact with Giovanni, her metadata also gets stored in the database of the ISB, possibly, for 30 years. In step five, the data collected by the providers and important companies, will then be processed and prepared for Hans. Afterwards, the CSB will grant Hans access to the requested data in a sixth step. Once the criminal prosecutor Hans has got access to the ISB, he begins to analyse all available data, depicted in step seven of Figure 5.2. Since there is no data protection law in case of surveillance, Hans is now allowed to do profiling and filtering even of sensitive and private data of Giovanni and his environment, which also includes the data of (innocent) Francesca. Because Hans needs to monitor Giovanni and his environment in real-time, he can order a so called "live-monitoring" from the CSB. He then is allowed to directly track

all data streams, once the affected telecommunication provider has established the setup for doing so. This live-monitoring is indicated in step eight of Figure 5.2.

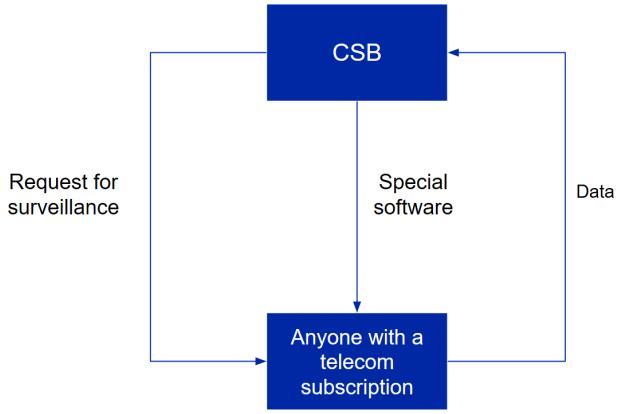


Figure 5.3: Process of surveillance part three.

In order to facilitate catching Giovanni, Hans orders Francesca to hand over the metadata stored on her communication devices. Since the previously applied suveillance possibilities have not provided sufficient evidence, Hans further orders the infiltration of Giovanni's and Francesca's computing devices with GovWare. All of the collected data is then stored in the database of the ISB (Figure 5.3). Once Hans has caught Giovanni and his gang, he stops the monitoring process. To finalize the process, the CSB has to contact the permission authority to inform them about the halt of the observation. The collected data about the criminal gang leader and his environment will be kept in the ISB database for up to 30 years.

5.3 Political Discussion

This section first discusses major points of criticism concerning the BÜPF are highlighted and brings them in relation to the ongoing process of its revision. Points highlighted in the first subsection called opponents to start a petition for a referendum against the revised legislation of the BÜPF, after it got passed in March 2016 [38]. Additional arguments of this committee are evaluated in the second subsection.

5.3.1 Criticism of the BÜPF

In February 2013, the Federal Council of Switzerland issued a legal gazette suggesting a total revision of the BÜPF [43]. Along with a broader ambit, new technologies should increase the possibilities for surveillance of electronic traffic. The Council argued, that the goal of this revision is not to increase the amount of surveillance, but to improve its quality, as it would become increasingly difficult to overcome surveillance along with the emerging technologies (like encrypted Internet telephony). To achieve this, the purview of the law needs to be extended and new surveillance technologies should be introduced. The Swiss society "Digitale Gesellschaft"[1] critically reflected the Federal Council's report to inform the public about the major legal changes and to question the feasibility of this revision [13]. As a result of the territorial principle, foreign telecommunication service providers which dominate the market would not be encompassed by the law. Although the Federal Council issued a clear statutory basis for the use of particular technical tools for surveillance (like IMSI-Catcher), their means are controversial. For example, the

[1] https://www.digitale-gesellschaft.ch

traffic of all devices within the signal reception of the IMSI-Catcher would get intercepted (c.f. Subsection 5.4.1). This way, the police could also monitor people with running mobile devices, who are simply located within the area of the suspect. This directly contradicts the goal of the revision, not to intercept more data.
In addition to the controversy of surveying technologies, the form and duration of storage from data collected related to surveillance activities is another major contentious point. In the legal gazette of the Federal Council, which constitutes the base for the BÜPF's legal reform, they proposed to increase the duration of data retention for metadata (e.g. name, address or birthday of a person, the respective subscribed services, contractual details etc.) from six to twelve months [43]. This extended data retention is heavily criticized by [13], as they miss the rationale, why extending the duration of data storage is an essential means for fighting crimes. The Federal Assembly justifies this extension with the fact, that data from a longer period is needed to fight criminality more effectively [43]. According to the law enforcement authorities, the period of six months often expired already before they were actually able to initiate the needed measure for surveillance. Providers however claim, that the longer data storage ultimately leads to higher costs incurred by acquisitions of more physical storage capacities, increased maintenance or additional retroactive activities [36, 43]. Moreover, external people draw attention to the fact that more data increases the endangerment when sensitive information gets in the wrong hands, constituting a higher risk of security [36].

5.3.2 Data Retention

According to the secrecy of telecommunication ("Fernmeldegeheimnis") regulated in Art. 13 Para. 1 of the Swiss Federal Constitution, each person has the right for attention to her postal and telecommunication traffic [39]. A retention of data like this would however be an intervention of the human civil rights, as no solid proportionality can be revealed [1, 13]. The Max-Planck institute further states that there is no evidence, that the data retention as practised in Germany and Switzerland could lead to more effective criminal investigations, and no ironclad proofs exist, that protective mechanisms would be weakened by the abolition of data retention [1]. These controversies about data retention and the fear of a failing revision or a referendum ultimately hindered the Federal Council to increase the duration for data storage and leave it legally fixed to six months as of the current version of the BÜPF [38].
When making use of stored data for an investigation of a criminal act, there is often the problem that large-scale surveillance programmes cannot initially help find evidence, moreover they help "connecting the dots" *after* a specific event has occurred [5]. This data-mining approach of retrofitting evidence into a case after having exercised undue surveillance may lead to contrary effects, disturbing the process of criminal investigation rather than accelerating it. In 2014, US president Barack Obama commissioned an independent oversight board to investigate the justification of the telephone records program defined in Section 215 of the American "Patriot Act" [28]. The evaluation concluded that the long-standing surveillance of telephony data proved to be useless in fighting terrorism or criminal activity. They could not find a single case in which such a surveillance program could help uncover or prevent terrorist attacks. This report raises doubts that intercepting telephone calls in general is helpful in investigations of larger national threats. Nevertheless, there was no indication of its usefulness in investigation of minor criminal offences.

5.3.3 Failed Referendum against the new BÜPF in July 2016

On the 18th of March 2016, the Federal Assembly passed the legislation about the revision of the BÜPF [38], based on the legal gazette published in February 2013 [43]. A few days after it got accepted, a coalition of political parties and organizations of the civil society launched a petition for a referendum against this new legislation.
Although the Federal Council proactively weakened the bill to prevent critic parties of launching a referendum, the opponents of the new legislation still heavily criticize its content. The extended area of application suddenly also affects small businesses or even groupings like living communities or families with shared Internet access, as they fall in the category of "provider of internal (tele-)communication networks" [36]. This leads to substantial additional costs, as these people are now conflicted with duties like providing metadata, endure surveillance or being reachable for the CSB if the they order so. Nevertheless, the chance for those people to actually be directly affected is very slight, since the majority of them are derived from large network providers. In this case, the CSB would assign a possible surveillance job to the party with the least technical effort for doing so, which predominantly are the large provider [38].
As of July 7th, the referendum narrowly failed [36]. The referendum committee could collect enough signatures, but their attestations by the residential communities could not be made in time [30, 36].

5.4 Technical Implications of the BÜPF

With the new version of the BÜPF being in force from January 2018 on, there are new requirements for telecommunication providers and some companies, chosen by the Federal Council. According to Section 7 in the new BÜPF, those firms have to be ready for surveillance all the time [38]. This constant readiness requires additional material and human resources for all providers and companies. Although it is not clearly defined, what this governmental readiness standard contains, it is clear, that there are hardware or software investments necessairy, independent of the party which has to come up for them [24, 33]. These additional resources are discussed in the following subsections.

5.4.1 More Data implies more Hardware

The new BÜPF reveals, that the Swiss Federal Council, the Federal Assembly and the criminal prosecution authorities plan to generally generate more data on the one hand, and to deregulate the legal boundaries for new surveillance techniques on the other hand [38]. Some of these new surveillance techniques were already used by the criminal prosecution authorities in the past years, namely antenna scans [21] or IMSI-Catchers [31].
An antenna scan (in german called "Antennensuchlauf" or "Digitale Rasterfahndung") is a technique, in which a telecommunication provider filters the required metadata, in order to find out, which subscriber was located around a specific antenna of the provider at a certain point of time. This is however not a real-time surveillance method. The technique serves the cause that the responsible authority is able to identify people they are looking for retrospectively. With the next version of the BÜPF, the government specifically allows the use of this technique [38]. Therefore it is valid to assume, that authorities will use this technique more frequently in the future.
IMSI (International Mobile Subscriber Identity) Catchers are devices which allow tapping and locating mobile phones [31, 35]. Such devices work as a middleware between a real antenna (base station) and its subscribers. Since mobile phones are optimized to gain highest quality of reception, they log into the base station with the strongest signal received. An IMSI-Catcher masquerades as a real base station and manipulates mobile phones within

a certain radius, to log in to them instead of the real base station. Additionally, because IMSI-Catchers are able to overcome the physical and SIM (Subscriber Identity Module) protection mechanisms, they can monitor phone calls, Internet traffic and other kinds of communication supported by the telecommunication providers network. Such devices therefore allow authorities to monitor mobile phone traffic in real-time. Alike antenna scans, the explicit allowance of IMSI-Catchers will probably lead to an increased amount of monitoring requests with these devices. Considering the proof of readiness in the new version of the BÜPF, such a development might require the acquisition of IMSI-Catchers either for TSPs or for prosecution authorities [38].
Although the data retention period of half a year has not been increased by the new BÜPF [37, 38], the new surveillance techniques and the possibility of new kinds of metadata determined by the Swiss Federal Council, will require more storage space for Internet providers, some specific companies, the CSB and the Swiss criminal prosecution authorities. This development will generally require more storage hardware space.

5.4.2 Need of new Software for Surveillance

According to Section 2 of the new BÜPF's specification [38], there will be a new IT-Service (ISB), which includes a database system and the ability to filter and process the data, supplied by the Telecommunication Service Providers (TSP) or certain companies chosen by the Federal Council. This software system also needs to have application programming interfaces (API), which allow TSPs to directly supply data to the ordering authority by granting them direct access to specific data. An important point in the specification of the ISB constitutes the adequate security of the used systems together with its contained data. Since there was neither a centralized IT-Service, nor a centralized IT-System for the purpose of surveillance in the BÜPF so far, the CSB will have to buy or establish at least a database system, a software for processing and filtering data, as well as APIs, which allow the operations mentioned above. Additionally, storing the collected data is challenging, because the data needs to be stored securely up to 30 years.
The upcoming version of the BÜPF specifies the standard behaviour of TSPs and some companies chosen by the Federal Council in case of surveillance. As mentioned in Subsection 5.2.4, the companies affected by the new BÜPF need to be prepared for surveillance at any time. All TSPs and companies chosen by the Federal Council have to obey the API defined by the ISB. With the new BÜPF, TSPs will need to adapt this API.
The Federal Assembly and the Federal Council have explicitly allowed the use of special software if the previous actions of surveillance were not successful or futile [38]. Such special software is also referred by the Federal Council as "GovWare" (abbreviation of Governmental Software). GovWare is specified as a piece of software, which is able to infiltrate a computing device to log what the user of that device is doing or to grant access to the files of the machine and its attached devices. Such a device could be a webcam or a microphone. Therefore, the new BÜPF basically allows the criminal prosecution authorities the use of lawful trojan horses. It is an open question who is going to develop this software and who exactly is going to use it. Additionally, there is the question whether to buy already existing GovWare or to make GovWare themselves. In both cases there are major challenges in adjusting the software to the needs of the authorities and the laws, which they have to comply with. Further, GovWare also needs to be maintained. In order to install governmental trojan horses on the computers and telephones, safety gaps in corresponding systems and programs (e.g. Microsoft Word, Adobe Acrobat, etc.) that are not known to the related software developers could be exploited [4]. For Example, in the USA, the National Security Agency (NSA) actually actively buys and hoards such security leaks, instead of informing the software companies about these issues [11]. Thus, as soon as the authorities would inject the malware, they would have no incentives to report these

safety gaps. Therefore, the authorities implicitly would accept that computers of all the citizens and firms stay unsafe.

5.5 Economic Implications of the BÜPF

As of beginning 2012, TSPs and Postal Service Providers (PSP) had to bear their investment costs for monitoring on their own, while operational costs were compensated by the Federal Council [38]. With the decision on the 9th of March 2012 regarding the total revision of the law being in place, the ISC-EJPD ("Informatik Service Center of the EJPD") and the CSB ordered KMPG Switzerland AG (KPMG) to estimate and analyse the costs that TSPs and PSPs faced due to implementation of the monitoring as required by the BÜPF. The final report is dated June 12th 2012 and is taken as a baseline for this report's analysis of economic implications of the BÜPF currently in force, summarized in the following Subsection [24]. A more recent article from Digitale Gesellschaft reveals costs implications for the new version of the BÜPF, becoming effective in January 2018 [12].

5.5.1 KPMG Report

For the estimation, KPMG developed a questionnaire and with the consent of ISC-EJPD, they sent it to a group of selected TSPs and PSPs. Additionally telecommunication and postal service providers were requested to provide the related data and were invited to interviews. For the sake of a better understanding, the given set of the companies were grouped into four clusters, i.e. big, medium, small TSPs and PSPs, correspondingly. The report was drawn on the basis of the information of eleven TSPs (four big, three medium, three small) and three PSPs. The clusters differ from each other in terms of the average size of their constituents, which is directly related to the portfolios they offer, and in the companies' network control capacities. The financial information related to the size of the clusters is presented in Table 5.1:

Service Provider Cluster	Sales (in 1'000 CHF)
Big	3'961'619
Medium	47'417
Small	14'128
Post	4'386'841

Table 5.1: Desriptive financial sales of TSP and PSP clusters [24]

All four big TSPs have their own telecommunication network and offer a large portfolio of telecommunication services, where three of them are so-called "Full Service Providers". The medium cluster differs from the big cluster naturally in terms of size, e.g. one of them provides solely mobile prepaid services. Small TSPs are rather young and growing companies that specialize on special niches of the telecom sector. In comparison to other clusters, big TSPs have the most substantial costs as they are able to offer a broader scope of monitoring.
The questionnaire developed by KPMG gathered all related information regarding the following yearly costs from 2007 to 2012 (where the numbers for the year 2012 were own companies' estimates): The scale of charges that consisted of personnel costs, expenditures on material, number of all required monitoring requests, investment costs, and the most important financial operating numbers.
Operational costs together with investment costs build the total cost structure, which is listed in Table 5.2. Although the numbers for 2012 represent solely expected costs, they

illustrate that the companies were already anticipating a substantial increase of costs, the new BÜPF would incur.

	Total Costs				Average Costs			
Year/Cluster	Big	Medium	Small	Post	Big	Medium	Small	Post
2010	21'029	337	288	10	5'322	94	96	3
2011	21'134	1'680	374	10	5'408	431	125	3
2012	**29'216**	**3'265**	**369**	**10**	**7'597**	**826**	**127**	**3**

Table 5.2: Costs Overview of TSP and PSP clusters (in 1'000 CHF), based on Tables 2, 3, 4, 5, 7, 10, 14, 16, 19, 22, 24, 25, 27, and 29 in [24]

The division of these total costs into operational and investment costs is presented in Table 5.3. A striking feature of this distribution is the fact, that the total costs of medium TSPs are comprised mostly of investment costs. Of them, however, only a small part is directly related to surveillance, which is completely the opposite situation for small TSPs.

	Operations				Investments			
Year/Cluster	Big	Medium	Small	Post	Big	Medium	Small	Post
2010	96%	35%	83%	100%	4%	65%	17%	0%
2011	93%	8%	87%	100%	7%	92%	13%	0%
2012	88%	4%	91%	100%	12%	96%	9%	0%

Table 5.3: Distribution of Total Costs of TSP and PSP clusters based on Tables 2, 3, 4, 5, 7, 10, 14, 16, 19, 22, 24, 25, 27, and 29 in [24]

5.5.1.1 Operational Cost Structure

As of 2011, the yearly operational costs of big TSPs were on the level of CHF 18.8 million, while medium and small TSPs each had to pay around CHF 0.2 million. The postal services companies did not incur any substantial operational costs related to monitoring. This distribution of costs takes place mainly due to the reason that big TSPs received a much larger number of monitoring requests. This is why they had to create and manage their own monitoring departments, while this task was still carried out along the usual operational duties for medium and small TSPs and PSPs. However, for small TSPs these operational costs were still very substantial, taking into account their size.

Three of four big TSPs expected increasing costs for the coming years on the basis of the additional requirements stated in the revised version of the BÜPF. One of the big TSPs argued, that one should expect an increase in the number of monitoring requests (especially with regard to IP monitoring), as well as in the variety of new technologies which allow monitoring. This would lead to a considerable increase in the investment, maintenance, and licensing costs.

One of the TSPs, whose maintenance and services of the network was provided by an external service provider pointed out, that, according to its own estimates, the costs that are related to the implementation of monitoring for this company are actually higher. This is caused by the fact that all important activities with the external service provider they rely on, have to be coordinated, which accounts for costs occasionally getting doubled.

On an absolute scale, the number of monitoring requests and their associated costs faced by medium and small TSPs, are not substantial in comparison to those of the big TSPs. However, proportional to the size of the respective enterprises, they are highly significant. For the cluster of the small TSPs, the business models vary to a great extent, which leads to differences in the types of monitoring activities they execute. This has a noticeable effect on the costs evaluation. As such, for two of three small TSPs, mainly the fixed costs matter, while the third small TSP exclusively faces variable costs.

What concerns the PSPs, the monitoring activities can be carried out by the employees as a part of their normal operation. There are not many monitoring requests (and costs related to them) that PSPs have to face. This applies also for the investment costs, which are not essential in this relation.

5.5.1.2 Investment Cost Structure

Regarding the investment costs related to surveillance in the year 2011, big TSPs invested CHF 1.5 million in total, while medium TSPs accounted for CHF 0.15 million. Small TSPs as well as PSPs did not report substantial investment costs related to surveillance in that year.

Year	Average	Total
2007	98'721	197'441
2008	100'277	300'832
2009	114'791	344'372
2010	260'138	780'414
2011	499'397	1'498'192
2012	**1'171'988**	**3'515'964**

Table 5.4: Investment costs of big TSPs (in CHF); presented numbers are derived from Tables 8, 20, 28, and 30 in [24]

For big TSPs, as Table 5.4 reveals, the investments transacted inside the clusters were increasing from year to year. Especially noticeable are the years 2011 and 2012: for three out of four big TSPs, the coming years bring much higher investment costs due to required additional monitoring activities that were expected after the total revision of BÜPF. Particularly what concerns the potential rise of Internet Protocol (IP) address tracings. However, as two big TSPs indicated, investment costs could potentially be used for automation of monitoring activities, which in its turn could lower future operational costs, as personnel costs would decrease. Nevertheless, since the investment costs are not compensated at all, it would be hard to implement this idea. Thus, the operational costs will not diminish and therefore stay the same. Only a small part of the investment costs for medium TSPs is related to monitoring activities and predominantly consist of those related to the everyday operations. The investment costs of small TSPs are substantial, both for the monitoring activities, and related to the everyday operations. Finally, investment costs for PSPs are not essential in terms of surveillance practices.
As the authors of the KPMG report highlight themselves, the numbers presented in the preceding subsections have to be taken into consideration carefully. These costs are calculated on sole estimations made by the respective service providers, based on different factors and a lot of assumptions [24].

5.5.2 Estimation of Costs for the latest Revision of the BÜPF

In 2014, the charitable organization Digitale Gesellschaft presented their own cost estimates for TSPs and PSPs, based on data they gathered from providers of different size [12]. Since there are no reliable sources of the presented estimation available, one can assume that the numbers presented in their report are biased, because Digitale Gesellschaft does not support the new BÜPF as one can readily verify on the official website of the anti-BÜPF campaign[2]. The reproduced cost numbers are depicted in Table 5.5.
It is of interest to compare the results from the KPMG report and the more up-to-date estimates presented by Digitale Gesellschaft. Comparing the results from Table 5.2 and

[2]http://www.stopbuepf.ch

Who	Number	Initial	Operations	Initial total	Operations total
Big-4 Telecom	4	12'000	6'000	48'000	24'000
Big ISPs	15	1'000	300	15'000	4'500
Medium ISPs/ Software providers	130	600	200	78'000	26'000
Small providers/ Firms and Institutions	1'800	90	40	162'000	72'000
Total	**1'949**	**13'690**	**6'540**	**303'000**	**126'500**

Table 5.5: The costs of the BÜPF in full operation (in 1'000 CHF) [12]

Table 5.5, one can see that the latter has divided the different companies into finer clusters. For a better comparison, this report merges the 130 medium ISPs/software providers and the 1'800 small providers/firms and institutions from the analysis done by Digitale Gesellschaft together in one cluster. This way, there are again three different clusters of TSPs, as it was the case in the report of KPMG (c.f. Table 5.1). The estimates for these newly created pseudo-small firms are presented in Table 5.6.
The central message of the article from Digitale Gesellschaft is, that the monitoring process is estimated to cost CHF 430 million for all providers in the first two years [12]. In the respective first year (2017), the TSPs have to introduce the necessary services for the new requirements posed by the BÜPF, and in the following year, they will face increased operational costs. Table 5.6 illustrates that the upcoming version of the BÜPF will impose much larger costs. On average the increase of costs concerns the big firms. In general, however, most of the total costs (i.e. CHF 338 million out of CHF 430 million) stem out from the big amount of small TSPs. This estimate is much higher than in the KMPG report, since the number of TSPs in the respective cluster is much higher.

		Total Costs			Average Costs		
Source	Year/Cluster	Big	Medium	Small	Big	Medium	Small
KPMG	2010	21.03	0.34	0.29	5.32	0.94	0.10
	2011	21.13	1.68	0.37	5.41	0.43	0.13
	2012	29.22	3.27	0.37	7.60	0.83	0.13
Digitale Gesellschaft	**2014**	**72.00**	**19.50**	**338.00**	**18.00**	**1.30**	**0.18**

Table 5.6: Updated Cost Estimates of TSP and PSP clusters (in million CHF) based on [12, 24]

5.5.2.1 Compensation Rate

The real number of total costs caused by the BÜPF is not officially available [14, 15, 23]. Altogether in 2012, cantons paid the CSB CHF 14.5 million. This amount however only includes the fees and compensations of the monitoring means. Investment costs for the providers to build the respective infrastructure are not covered. The CSB compensated the providers with CHF 9.4 million (i.e. 64% of the total amount). These numbers stem from the evaluation of the openly-accessible statistics prepared by Digitale Gesellschaft in [23]. According to the evaluation of [23], the three big telecom providers at that time (Sunrise, Swisscom and Orange) were compensated with CHF 9.3 million (Sunrise: CHF 3.6 million, Swisscom: CHF 3.3 million, Orange: CHF 2.4 million). The rest of the amount (CHF 0.1 million) was distributed among the remaining TSPs. Swisscom and Orange confirmed the cost dimensions [23]. However, the compensation covered only about 40-50% of their actual operational and running costs. Moreover, Orange shared their opinion that monitoring of criminal prosecutors is clearly part of the tasks the Federal Council has to deal with. Sunrise did not comment the numbers. More up-to-date compensation

rates are equal to 64% as well [14, 15]. Whether this is a coincidence or an obvious trend, cannot be said. Regardless of the numerous requests by the media, the CSB refused to give any comments regarding concrete compensation numbers and their development [23]. They refers only to the scope of the regulation of fees.

5.5.3 Impact on Telecommunication Provider Market

In general, the new BÜPF will push existing small firms and start-ups away from the market [19, 29]. The high costs implied by the new BÜPF are going to ruin around 120 small providers since they are not able to provide the necessary 24-hours monitoring services. High investment costs arise since TSPs have to be ready to give access to the metadata of their users to the CSB. Possibly, they will have to cooperate by installing GovWare. Actually, not only the big TSPs are bound by the BÜPF, moreover also any software in which any kind of communication service is incorporated (e.g. bookkeeping software with the chat-function) falls under its force, and hence incurs these investment costs [19].
The new BÜPF may also lead to vanishing message applications business for local firms. For instance, Martin Blatter, co-founder of the messaging service *Threema*[3], explains, that storing metadata contradicts his philosophy and forces the company either to move abroad or to develop a smart technological solution against that, like Threema is doing [29].
From the economic analysis of costs that telecommunication providers face due to the BÜPF, additional inferences can be made. With the issues small providers face, these consequences might result in the oligopoly of big providers in the telecommunication market. In that case, the competition among the medium providers will escalate, which represents a threat for the small providers to die out completely. One would also anticipate that aspiring small providers and potentially interesting start-ups face higher entry barriers. As a result, the Swiss telecommunication market would have less opportunities for small businesses.

5.6 Social Implications of the BÜPF

This section analyses the impacts of the BÜPF from a social perspective. After taking a general viewpoint of surveillance in public, implications on privacy/security issues from data collected in surveillance activities are critically highlighted and questioned with regard to data misuse or violations of rights.

5.6.1 Panopticon Effect

Today, it is virtually impossible to hide surveillance when staying at public places like train stations, shopping malls or libraries. CCTVs monitoring the tracks, detectors preventing from stealing goods or just the swipe of the credit card when purchasing a coffee records and potentially store crumbs of your identity. This problem of being at the mercy of surveillance in public places is a widely disputed topic [1, 34, 36].
Swiss citizens have the right of protection of their privacy, as stated in Art.13 of the Swiss Federal Constitution and the "right of free (physical and psychological) movement" (Art. 10) [39]. The article of Slobogin illustrated that actions of humans being videotaped are influenced, based on several famous law cases [34]. These examples support the fact, that human behaviour is controlled to some degree by public surveillance. Even if the surveying methods used are legally allowed, public surveillance can infringe interests in locomotion

[3] https://threema.ch

"to a legally cognizable degree" [10, 34]. In Andrew Taslitz's article about privacy issues related to public surveillance, he describes privacy as "*a means of presenting to others only the parts of ourselves we want them to see*" [46]. Depending on the situation, people show different versions of themselves and act differently. Conscious about being monitored, people adjust their behaviour, which is also known as the effect of the Panopticon [3]. This effect has its origins in prisoners supervision. An observation tower for the jailers in the center of the prison, circled with the cells, conveys the feeling of total surveillance. Prisoners always feel monitored and therefore act as they would be watched all the time. This induced state of being monitored can therefore be used as a means of control and power, although only one single watchman might be present.
Increased surveillance and data retention can however also lead to increased security. The panoptic effect also leads to different behaviour patterns for criminals [32]. Electronic means of communication may be reduced to a minimum, complicating the activities of criminal networks. Concrete evidence that intensified surveillance leads to more security in Switzerland however barely exists, as the Federal Statistical Office does not collect data of procedures used in crime detection [17].
Criminal prosecution authorities in Switzerland can request surveillance of people while dealing with criminal felonies in accordance with Art. 269 of the code of criminal procedure [41]. Having a closer look at this article reveals, that also minor offences are listed in there (e.g. minor forms of property damage, depiction of violence or usage of doping). Such actions grant permission to the authorities to access stored metadata of the suspected person and the people in their environment, stored by the postal or telecommunication providers, as prescribed by the BÜPF. People therefore become increasingly more aware about what information they are sharing, with whom, in which way, and what words they are using. Even if people are communicating over the Internet and only metadata is collected, it might be possible to draw inferences from them about the conversation's content. The same might also apply if one of the communication partners is under surveillance, while the other one is not involved in the criminal offence at all.

5.6.2 Can we trust the Government?

The circumstances under which the use of special programs for surveillance, like IMSI-Catcher or GovWare is justified, comply with the new revision of March 2016 [43]. Nevertheless, as [13] shows, data of non-involved persons may also be collected in the process of monitoring a suspicious person. Besides privacy related concerns, being (unwillingly) at the mercy of surveillance in public places, also raises questions of trust and security. Can people still trust the government/national intelligence services that their actions comply with the law? What about abuse of collected data? The European Convention on Human Rights (ECHR) guarantees everyone the right of privacy even if their country is not part of the EU. The problem is however if data of European citizens gets processed by countries outside Europe. Exemplary, if a US intelligent service processes this data, it is not protected the same way as if it belonged to an US citizen [5]. A common regulation on privacy rights to data or an international treaty specifying a bill with the US, having the biggest secret service of the world, is however not conceivable as they do not seem ready to accept any constraining changes which are not present in the US Constitution [5]. Evidence released by Edward Snowden reveals that data is exchanged between European intelligence services and the US [5, 25]. A possible reason for these actions might be to bypass different legal frameworks by "outsourcing" surveilling activities to countries with weaker oversight regimes, as it could be ascertained to some extend for the Government Communications Headquarters (GCHQ) of the United Kingdom and the NSA of the United States of America [5, 25]. Having regard to the American "Patriot Act", providers

of telecommunication services are then compelled to reveal their data to the authorities – without any judicial approval.
In 2011, the German (and Europe's largest) association of hackers, called "Chaos Computer Club", managed to hack the state trojan of Germany and revealed serious security flaws [9]. Not only that the server was stationed in the USA but also the fact that transferred data was very badly encrypted (server commands were even fully unencrypted) poses immediate threats to security. This way, unauthorized third parties could easily telecommand the software. The same firm, which programmed the German GovWare, also sold their software product to Switzerland. Official evidences, that Switzerland is involved in data exchange with the US intelligence services to possibly circumvent legal regulations, however, do not exist.
Recent terrorist incidents like the attacks in Paris (November 2015), suicide bombings in Brussels (March 2016) and Germany (July 2016), or the brutal assassination in Nizza (July 2016) deeply shook the people, including the Swiss citizens. The fact that terrorism is ubiquitous also sensitized Switzerland [20]. The danger of so-called "lone wolfs"[4] increases and presenting new challenges for the Swiss government. Certainly, this was also a reason, why the revision of the Swiss law for intelligence service (free translation of "Nachrichtendienstgesetz") was clearly accepted by the Swiss citizens [6]. The government was given more trust by letting them extend the possibilities for surveillance, which might also have a positive effect (in terms of public acceptance) on the BÜPF. People want the government to actively monitor criminals and prevent them from their actions. However, being monitored themselves (what might also be the case even they are innocent, as seen before) is mostly out of the question. Violations of human rights, privacy or data security remain the main points for criticism. Gaining people's trust by providing proof of successful surveillance activities, without endangering their privacy or security on the other side, is therefore a delicate task the government has to cope with.

5.7 Evaluation and Discussion of the BÜPF

This section illustrates important interconnections and mutual influences between the different impacts of the new BÜPF from an economic, social and technical perspective. First, the economic implications of the technical aspects are discussed. Afterwards, influences of these technical aspects on the society are examined. Finally, the outcomes caused by economic implications on the situation in society are listed. Additionally, related cases and frameworks for monitoring in other countries are discussed and compared to the situation in Switzerland.
Subsection 5.2.3 points out that one of the main requirements of the new BÜPF is the readiness of providers and any related parties in general, to supply the data in real-time. There are two main aspects to it: On the one hand, the necessity of more storage space arises for providers, as an outcome of more data which needs to be saved (i.e. the hardware aspect). The second aspect is the software aspect, which is on the other hand, directly related to the ISB. Acquisition of more hardware is linked to exploiting of antenna scans that enable retrospective monitoring, as well as more advanced devices such as IMSI-Catchers. This allows not only to monitor the subscribers, but also intercepting mobile traffic in real-time. Together, this creates even a higher need for storage space. Naturally, the purchase of additional hardware is inevitably linked to higher cost for TSPs. These costs are, as highlighted in Section 5.5, not compensated by the Federal Council and

[4] Lone wolfs or lone-wolf terrorists are single perpetrator committing terroristic acts on their own, without assitance or command of any (terroristic) group, however possibly driven by such ideologies. This fact of being isolated from bigger terroristic networks makes them extremly difficult to detect for intelligence agencies [2].

represent the main threat for small firms, being the potential reason of their complete vanishing from the market. Related operational costs will force the providers to rise their fees, which induces additional expenditures for their client base that is comprised of the commercial clients as well as ordinary citizens.

The software aspect is centered around the ISB database, which involves its maintenance, as well as development and installation of the API that represents the main transition tool for supplying this database with the data collected by TSPs. The database and additional software (e.g. GovWare) of high security, robustness, efficiency and secrecy will force the CSB to form new IT teams. On the one hand, for developing the corresponding software, on the other hand for interpretation and analysis of the collected data that can be ordered by criminal prosecutors. Additionally, these newly established teams will need further education and training in the future, taking into account that technological progress always moves ahead of any legal framework. This situation creates a need in labour and thus an employment opportunity for IT professionals.

Not only a significant economic, but also a strong social impact is implied by the technical aspects of the BÜPF. Most notably, this concerns data privacy. Since the society forms a strongly interconnected network, there is no certain guarantee for someone innocent not to get monitored as a part of a larger surveillance procedure. Uninvolved citizens will probably never find out whether they were monitored or not. The only eavesdropping-free method of communication remaining is the private conversation, though one might also cast doubt, since most of the digital devices have an integrated microphone which may be secretly listening to the conversation. Hence, even the luxury of a private conversation where the shared information might be worth nominally a large amount of money, is limited.

From a legal perspective, the EJPD stated that each case of surveillance of telecommunication constitutes an intervention in human rights [18]. Additionally, they claim that data retention also endangers the freedom of expression, as the trust of people in means of communications is affected which may lead to different communication behaviours. Nevertheless, high legal boundaries (in form of strong suspicion of a serious crime, along with a permit from the "Zwangsmassnahmengericht") justify these interventions. These circumstances further provoked discussions related to this manner, also outside the border of Switzerland [45, 48]. Germany, as well as Austria, both classified the data retention as unlawful interference in the basic rights already earlier. They significantly enhanced the requirements and regulations for data retention in 2010 (Germany) and 2014 (Austria), respectively [7, 8]. In case of Austria, data retention even got abolished. A constitutional jurisdiction (in German *Verfassungsgerichtlichkeit*) like in Austria, Germany or other European countries does not exist in Switzerland. All the federal laws are binding, even if they contradict the constitution [22]. As of December 2015, a new law regulating the storage of metadata for telecommunication providers in Germany got adopted. The fact, that the BÜPF got revised in the same time showcases the rethinking of importance of data, as it became inevitable to draw on data from different sources for tracking down persons of interest.

With respect to the case study presented in the introduction, the report pointed out arguments supporting and contradicting the intensified surveillance caused by the new BÜPF. On the one hand, it is necessary, that criminal enforcement authorities are equipped with state-of-the-art technologies to track down criminals [43, 44, 45]. The requirements for making use of these technologies are high – high enough to justify its application even if it is contradictory to other laws. Criminals should not have the possibility to get away without punishment because of loopholes in the law. Further, more surveillance may lead to an increase of the Panopticone effect for potential criminals, preventing them from turning criminal. On the other hand, the surveillance possibilities legalized by the new BÜPF, especially the GovWare, breaks the basic human rights for privacy [39, 34]. In

the case of Dr. Bob and Alice, it also hurts the medical secrecy [40]. Therefore, the new BÜPF may also be abolished by the ECHR. The new BÜPF grants the Swiss authorities more power over the people they serve. Hence, the potential damage caused by misuse or abuse of this power is higher. Additionally, data retention and its duration increase the risk of data leakages in the future.

5.8 Summary and Conclusion

In this report different impacts of the new Swiss law concerning the monitoring of postal and telecommunication traffic (BÜPF) were analyzed from different angles. The fast growing technology fundamentally changes the way of surveillance, as new methods and number-crunching possibilities are introduced. This also implies changes in the law, as physically eavesdropping the neighbour's door, has been replaced by sophisticated software, enabling people to "listen" to other's conversations from all over the world. The history of the BÜPF has shown, that the law, however, always lags behind the technological evolution, revealing possible loopholes for (solely) legal justification of surveillance activities. Another problem states the "passive surveillance" of uninvolved people, as illustratively seen with IMSI-Catchers. Without people's knowledge, data (e.g. in form of phone calls) may get intercepted and possibly stored for up to 30 years. The question, whether privacy of those people is violated or not cannot easily be answered. Too many factors play a role (e.g. under which circumstances is the suspect monitored? Where will the data be stored? Physically but also geographically? How is this data encrypted and/or protected against potential misuse?), making it more a matter of a perspective argumentation than a model answer. From a governmental perspective, the new version of the BÜPF does however not aim to increase the amount of surveillance, but to improve its quality, as it would become increasingly difficult to overcome surveillance along with the emerging technologies [43]. By legally providing the authorities more possibilities for surveillance, this statement is however barely feasible. More ways of surveillance automatically leads to more surveillance, as the suspicious people are just monitored with different means at the same time.

The new BÜPF requires the government and the companies affected by it to invest in new hardware and software. It is not clear yet, which investment costs will be carried by whom as the standard of surveillance is not yet defined by the Federal Council. Nevertheless, some of the Swiss cantons already invested in new software and hardware by buying IMSI-Catchers or GovWare. If telecommunication providers are ordered to store metadata of its customers, or surveillance of a person is assigned to them, different costs emerge. First of all, these companies (may) have to invest in such new technologies for monitoring purposes or data retention, but also operation (e.g. maintenance) is linked to costs. For the latter, the Federal Council compensates the provider, for any investment costs they have to come up themselves.

Taking the economic perspective into consideration, it is expected, that the new BÜPF will cause more costs for the government and the affected companies. Although the estimations differ in numbers, they all agree, that the small and medium size telecommunication companies will suffer the most. The new BÜPF may lead to a consolidation of the telecommunication market. Therefore, one can expect, that consumer prices will increase as a consequence.

From a social and psychological perspective, the report takes the effect of the panopticon into consideration, depicting that people behave differently with the knowledge of (possibly) being monitored. As a result, surveillance can lead to an automatic and centralized functioned control instance for the government.

Knowing all the various involved parties of the BÜPF, different perspectives could be taken which lead to the following conclusions. Having a look at the BÜPF's history its latest revision, one can see that there will always be a controversy of legal surveillance practices and the right of privacy. People want both, on the one hand, control and surveillance of criminals, and on the other hand, complete privacy. This is also a reason why the questions raised in the introduction of this report cannot be conclusively answered.
Although the Swiss government extended the possibilities of surveillance, by granting the authorities nearly every right to collect and store data, there are still strict rules for the application of the new BÜPF. Every case of surveillance needs to be justified by a judge. In comparision, Switzerland has overall higher boundaries for applying surveillance laws than many other countries (e.g. USA, UK) allowing lawful monitoring [5, 25, 38]. These regulations, however always lag behind the technological progress of surveillance, entailing legal loopholes. Such gaps in the law get exploited by different intelligence services – even beyond national borders.
With the revised BÜPF in action, it is expected that more (especially sensitive) data will be collected. As a result, the possible damage caused by an attack or data leak will be more severe than before. Guaranteeing data security will pose big challenges for the government, as in the case of a data leakage, they will loose the trust of the people.

5.9 Glossary

API	Application Programming Interface
BÜPF	Federal law concerning the monitoring of postal and telecommunication traffic (free translation of "Bundesgesetz betreffend der Überwachung des Post- und Fernmeldeverkehrs")
CCTV	Closed Circuit Television
CSB	Central Service for BÜPF (free translation of "Dienst zur Überwachung des Post- und Fernmeldeverkehrs")
ECHR	European Convention on Human Rights
EJPD	Federal department of Justice and Police (free translation of "Eidgenössisches Justiz- und Polizeidepartement"
EU	European Union
GCHQ	Government Communications Headquarters (Major Internet surveillance agency of the United Kingdom)
ICT	Information and Communications Technology
IMSI	International Mobile Subscriber Identity
IP	Internet Protocol
ISB	IT-Service BÜPF (free translation of "Informatiksystem zur Verarbeitung von Daten im Rahmen der Überwachung des Fernmeldeverkehrs")
ISC-EJPD	Informatics Service Center of EJPD
ISP	Internet Service Provider
NSA	National Security Agency (of United States of America)
PSP	Postal Service Provider
PTT	Post Telefon- und Telegrafenbetriebe (government owned company until 1992)
TSP	Telecommunication Service Provider
USA	United States of America

Bibliography

[1] Albrecht, H.J. et al.: *Schutzlücken durch Wegfall der Vorratsdatenspeicherung*; Report, Max-Planck-Institute, Department of foreign and international penology, July, 2011, `https://www.mpg.de/5000721/vorratsdatenspeicherung.pdf`, Last Accessed: November, 17, 2016.

[2] Bakker, E. and De Graaf, B.: *Lone Wolves: How to Prevent This Phenomenon?*; Article, International Center for Counter-Terrorism, The Hague, Netherlands, November, 2010; `https://www.icct.nl/download/file/ICCT-Bakker-deGraaf-EM-Paper-Lone-Wolves.pdf`, Last Accessed: December, 1, 2016.

[3] Bentham, J. and Bowring, J.: *The Works of Jeremy Bentham*; Edinburg, W. Tait, 1843

[4] Berger, G., Schweizer Radio und Fernsehen: *Das will das neue Büpf: Daten länger speichern und Staatstrojaner*; Article, March, 11, 2014. `http://www.srf.ch/radio-srf-3/digital/das-will-das-neue-buepf-daten-laenger-speichern-und-staatstrojaner`, Last Accessed: November, 17, 2016.

[5] Bigo, D. et al., Centre for European Policy studying: *Mass Surveillance of Personal Data by EU Member States and its Compatibility with EU Law*; Report, Brussels, November, 2013, `http://tinyurl.com/hhml7wj`, Last accessed: November, 17, 2016.

[6] Bundesgesetz über den Nachrichtendienst; `https://www.admin.ch/gov/de/start/dokumentation/abstimmungen/20160925/nachrichtendienstgesetz.html`; Last Accessed: November, 17, 2016.

[7] Bundeskanzleramt Österreich; `https://www.help.gv.at/Portal.Node/hlpd/public/content/246/Seite.2460406.html`; Last Accessed: December 12, 2016.

[8] Bundesverfassungsgericht Deutschland; *Konkrete Ausgestaltung der Vorratsdatenspeicherung nicht verfassungsgemäß*; Media Release, March, 02, 2010, `http://www.bundesverfassungsgericht.de/pressemitteilungen/bvg10-011`; Last Accessed: December 12, 2016.

[9] Chaos Computer Club: *Chaos Computer Club analysiert Staatstrojaner*; Article, August, 2010, `http://www.ccc.de/de/updates/2011/staatstrojaner`, Last Accessed: November, 17, 2016.

[10] Denninger, E.: *Das Recht Auf Informationelle und Innere Sicherheit: Folgerungen aus dem Volkszählungsgesetzurteil des Bundesverfassungsgerichts*; Print, Kritische Justiz, Vol. 18, No. 3, 1985, pp. 215-244, `http://www.jstor.org/stable/23996361`, Last Accessed: November, 17, 2016.

[11] Digitale Gesellschaft Schweiz: *Massenüberwachung: Zugriffsmöglichkeiten auf Schweizer Datenkommunikation im Ausland*; Article, February, 16, 2015, http://tinyurl.com/zr998xx, Last Accessed: December, 01, 2016.

[12] Digitale Gesellschaft Schweiz: *Neues Überwachungsgesetz BÜPF würde die Wirtschaft in den ersten zwei Jahren 430'000'000.- kosten - und 120 Firmen ruinieren*; Article, September, 23, 2014, http://tinyurl.com/gtxdrzw, Last Accessed: November, 17, 2016.

[13] Digitale Gesellschaft Schweiz: *Kritik an der BÜPF-Revision*; Article, April, 27, 2013, https://www.digitale-gesellschaft.ch/uploads/2013/04/kritik_20130427.pdf, Last Accessed: November, 17, 2016.

[14] Digitale Gesellschaft Schweiz: *Swiss Lawful Interception Report 2015*; Report, March, 2, 2015, https://www.digitale-gesellschaft.ch/uploads/2015/03/SLIR_2015.pdf, Last Accessed: November, 17, 2016.

[15] Digitale Gesellschaft Schweiz: *Swiss Lawful Interception Report 2016*; Report, March, 3, 2016, https://www.digitale-gesellschaft.ch/uploads/2016/03/SLIR_2016.pdf, Last Accessed: November, 17, 2016.

[16] Digitale Gesellschaft Schweiz: *Überwachungsgesetz BÜPF: Mit Staatstrojanern auch gegen Bagatelldelikte*; Article, March, 18, 2016, https://www.digitale-gesellschaft.ch/2016/03/18/mit-staatstrojanern-gegen-bagatelldelikte/, Last Accessed: November, 17, 2016.

[17] Digitale Gesellschaft Schweiz: *Wo bleiben die Fakten?*; Report, July, 19, 2011, https://www.digitale-gesellschaft.ch/2011/07/19/wo-bleiben-die-fakten/, Last Accessed: November, 17, 2016.

[18] EJPD: *Fernmeldeüberwachung: Hohe gesetzliche Hürden schützen Grundrechte*; Media Release, July, 01, 2014, http://www.ejpd.admin.ch/ejpd/de/home/aktuell/news/2014/2014-07-01.html, Last Accessed: December, 12, 2016.

[19] Faki, S., Toggenburger Tagblatt: *Der Lauschangriff ruiniert 120 Firmen*; Article, June, 14, 2015. http://www.toggenburgertagblatt.ch/ostschweiz-am-sonntag/politik-und-wirtschaft/Der-Lauschangriff-ruiniert-120-Firmen;art304159,4258211, Last Accessed: November, 17, 2016.

[20] Federal Department of Foreign Affairs: *Terrorismusbekämpfung*; Article, September, 2016, https://www.eda.admin.ch/eda/de/home/aussenpolitik/sicherheitspolitik/neue-sicherheitspolitischeherausforderungen/terrorismusbekaempfung.html, Last Accessed: November, 17, 2016.

[21] Fonjallaz, J. et al.: *BGE 1B 376/2011, Überwachung des Post- und Fernmeldeverkehrs; Antennensuchlauf, Rasterfahndung - Beschwerde gegen die Verfügung vom 12. Juli 2011 des Zwangsmassnahmengerichts des Kantons Aargau*; Judgement of the federal court, November, 3, 2011.

[22] Haller, w., Kölz, A. and Gächter, T.: *Allgemeines Staatsrecht* 1st Edition, 2013. Zürich: Schultess.

[23] Hanimann, C., Die Wochenzeitung: *Die Kosten der Überwachung*; Article, March, 20, 2014, http://www.woz.ch/1412/ueberwachungsgesetz-buepf/die-kosten-der-ueberwachung, Last Accessed: November, 17, 2016.

[24] Haymoz, A. et al., KPMG AG: *Bericht "Erhebung und Analyse der Kosten der Post- und Fernmeldeüberwachung"*; Technical Report, Bern, June, 12, 2012, https://www.bj.admin.ch/dam/data/bj/sicherheit/gesetzgebung/fernmeldeueberwachung/ber-isc-ejpd-fda-pda-d.pdf, Last Accessed: November, 17, 2016.

[25] Hopkins, N. and Ackermann, S., The Guardian: *Flexible laws and weak oversight give GCHQ room for manoeuvre*; Article, August, 2, 2013, https://www.theguardian.com/uk-news/2013/aug/02/gchq-laws-oversight-nsa, Last Accessed: November, 17, 2016.

[26] Hostettler, O. and Klee, M., Beobachter: *Liberalisierung - Der Frust mit der Post*; Article, April, 2, 2010, http://www.beobachter.ch/justiz-behoerde/buerger-verwaltung/artikel/liberalisierung_der-frust-mit-der-post/, Last accessed: November, 17, 2016.

[27] Klaus, S., Mathys, R.: *"The Best of BÜPF" - Was ändert sich mit der Revision?*, Report, September, 22, 2016, http://tinyurl.com/gpxsze2, Last Accessed: December, 12, 2016.

[28] Medine, D. et al.: *Report on the Telephone Records Program Conducted under Section 215 of the USA PATRIOT Act and on the Operations of the Foreign Intelligence Surveillance Court*; Report, January, 23, 2014, https://www.pclob.gov/library/215-Report_on_the_Telephone_Records_Program.pdf, Last Accessed: November, 17, 2016.

[29] Metzler, M., Neue Zürcher Zeitung: *Keine Abstimmung über Trojaner*; Article. http://www.nzz.ch/nzzas/nzz-am-sonntag/buepf-massenueberwachung-gefaehrdet-schweizer-firmen-ld.88792, Last Accessed: November, 17, 2016.

[30] Referendum BÜPF - Homepage of the referendum platform: https://www.buepf.ch/; Last Accessed: December, 9, 2016.

[31] Rothenberger, J., Tagesanzeiger: *Swico zum Referendum gegen das BÜPF*; Article, March, 14, 2013, http://www.tagesanzeiger.ch/digital/daten/Das-Arsenal-der-Ueberwacher/story/16810462, Last Accessed: November, 17, 2016.

[32] Rudhayaini, V. M.: *Knowledge is Power: The Internet Panopticon as a Weapon against Terror*; Report, School of Oriental and African Studies, University of London, May, 19, 2016, http://www.e-ir.info/2016/05/19/knowledge-is-power-the-internet-panopticon-as-a-weapon-against-terror/, Last Accessed: November, 17, 2016.

[33] Schweizerischer Wirtschaftsverband der Informations-, Kommunikations- und Organisationstechnik: *Swico zum Referendum gegen das BÜPF*; Article, March, 18, 2016, http://www.swico.ch/downloads/dokumente/stellungnahme-von-swico-zum-buepf-referendumpdf/3329, Last Accessed: November, 17, 2016.

[34] Slobogin, C.: *Public Privacy: Camera surveillance of public places and the right to anonymity*; Mississippi Law Journal, Vol.72, 2002, pp. 213-299, http://dx.doi.org/10.2139/ssrn.364600, Last Accessed: November, 17, 2016.

[35] Sokolov, D., c't Magazin für Computertechnik: *Digitale Selbstverteidigung mit dem IMSI-Catcher-Catcher*; Article, August, 27, 2014, http://www.heise.de/ct/artikel/Digitale-Selbstverteidigung-mit-dem-IMSI-Catcher-Catcher-2303215.html, Last accessed: November, 17, 2016

[36] Stop BÜPF - Official homepage of the BÜPF referendum committee: https://stopbuepf.ch/; Last Accessed: November, 17, 2016.

[37] Swiss Confederation: *Bundesgesetz betreffend die Überwachung des Post- und Fernmeldeverkehrs*; Legal Text, October 2000, https://www.admin.ch/opc/de/classified-compilation/20002162/index.html, Last Accessed: November, 17, 2016.

[38] Swiss Confederation: *Bundesgesetz betreffend die Überwachung des Post- und Fernmeldeverkehrs*; Legal Text, March 2016, https://www.admin.ch/opc/de/federal-gazette/2016/1991.pdf, Last Accessed: November, 17, 2016.

[39] Swiss Confederation: *Federal Constitution of the Swiss Confederation*; Legal Text, January, 1, 2016, https://www.admin.ch/opc/en/classified-compilation/19995395/201601010000/101.pdf, Last Accessed: November, 17, 2016.

[40] Swiss Confederation - Homepage of the medical secrecy: https://www.edoeb.admin.ch/datenschutz/00768/00808/00831/index.html?lang=de; Last Accessed: December, 14, 2016.

[41] Swiss Confederation: *Schweizerische Strafprozessordnung (StPO)*; Legal Text, October, 1, 2016, https://www.admin.ch/opc/de/classified-compilation/20052319/201610010000/312.0.pdf, Last Accessed: November, 17, 2016.

[42] Swiss Confederation: *Volksabstimmung über das Nachrichtendienstgesetz (NDG)*; Legal Text, September, 2, 2016, http://www.vbs.admin.ch/de/themen/nachrichtenbeschaffung/nachrichtendienstgesetz.detail.news.html/wissenswertes/2016/160902.html, Last Accessed: November, 17, 2016.

[43] Swiss Federal Council: *Botschaft zum Bundesgesetz betreffend die Überwachung des Post- und Fernmeldeverkehrs* ; Federal Gazette, February, 27, 2013, https://www.admin.ch/opc/de/federal-gazette/2013/2683.pdf, Last Accessed: November, 17, 2016.

[44] Swiss Parliament: *Bundesgesetz betreffend die Überwachung des Post- und Fernmeldeverkehrs. Änderung*; Debate Report, Bern, March, 10, 2014, https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20130025, Last Accessed: November, 17, 2016.

[45] Swiss Parliament: *Telefonüberwachung - Ständerat gegen längere Vorratsdatenspeicherung*; Debate Report, Bern, December, 7, 2015, https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/legislaturrueckblick?AffairId=20130025, Last Accessed: November, 17, 2016.

[46] Taslitz, A. E.: *The Forth Amendment in the Twenty-First Century: Technology, Privacy, and Human Emotions*; Law and Contemporary Problems, pp. 125-188, 2002, http://scholarship.law.duke.edu/lcp/vol65/iss2/7/, Last Accessed: November, 17, 2016.

[47] Thommen, C., grundrechte.ch: *Wer BÜPFt mich denn da?*; Article, September, 26, 2014, `https://www.grundrechte.ch/wer-buepft-mich-denn-da.html`, Last Accessed: November, 17, 2016.

[48] Verfassungsgerichtshof (VFGH) Österreich: *Gesetze zur Vorratsdatenspeicherung in Österreich verfassungswidrig*; Media Release, June, 27, 2014, `https://www.vfgh.gv.at/downloads/presseinformation_verkuendung_vorratsdaten.pdf`, Last Accessed: December, 12, 2016.

Chapter 6

The Breaching of Cloud SLAs: Why it Happens and How to Prove it?

Lukas Eisenring and Catrin Loch

Cloud Computing has become a well-known technology over the last few years, which is said to have an revolutionary impact on the information industry and IT industry [13, 52]. Mobile Cloud Computing, for example, is used in everyday activities such as checking facebook, emails, documents which are shared with friends or co-workers and many more. There exist a lot of advantages such as accessing data from everywhere, but also disavantages such as consuming a large amount of power [1, 54] or the breaching of cloud service level agreements (SLA).
The aim of this report is to analyse the breaching of cloud SLAs with respect to the two questions: Why it happens and how to prove it?
"Breaching of cloud SLAs" is understood as a company not providing the quality of service, which has been defined in the SLA and agreed upon, to a customer. To discuss the given research problem virtualisation and monitoring are presented.

Contents

6.1 Introduction

Many advances in information and communication technologies have been transforming the world, such as Cloud Computing (CC), wireless communication and competitive mobile device industries [54]. Overall CC has increased fast in the past decade and is now essential for the big computing resources to make the modern web available [20]. Also the virtualisation systems have gained a lot in efficiency and popularity with the enormous demand in increasing computing power [28, 29]. A large hype about the cloud, which was created by marketing, led to a strong user expectation pressure and expectations towards its characteristics, which partially could not be fulfilled [52]. Therefore CC has not only positive effects on the computing resources, but also negative ones, such as the breaching of cloud service level agreements (SLA). The aim of this report is to examine why breaching of cloud SLAs happen and how to prove it.
First of all, the purpose of SLAs is to formally define the parameters of service the provider guarantees to deliver. Concluding that, after agreeing upon an SLA, the specified parameters are monitored in order to detect agreement breaches and to verify their correctness against the promised quality of service (QoS). On the other hand, the user needs to be able to ensure, that the provided monitoring information is the right data and to interpret the terms with respect to the promised quality. Furthermore, when a SLA breach is detected, an appropriate remedy procedure (also defined in the contract) is applied. In the end, identifying the violations and calculating penalties might prove quite challenging and QoS monitoring and management are, in the end, old concerns in terms of IT research and development [34, 52].
Finally, the given research question is only discussed in the context of public clouds.

The structure of this report is as follows. First, the theoretical background of CC is presented. Afterwards the participating entities in the analysed problem are discussed. The next part of the report describes the problems and solutions of virtualisation and its components such as the memory and network access. Then monitoring is reviewed with utilisation of different examples, such as Google Stackdriver Monitoring or Kaseya Traverse. The last part of this report consists of the discussion of the solutions and a short summary including some suggestions.

6.2 Theoretical Background

Overall the concept of CC is not a novelty in itself, as it is thought by a lot of users. In fact, the priniciples are said to be arosen from a direct industrial need to improve resource utilisation without impacting on consumer requirements, i.e. use the available resources more efficiently [52]. More specific CC is seen as the developing result of grid computing, distributed computing, parallel computing, utility computing, network storage technologies, virtualisation etc. or shortly traditional computer technology [13]. As CC is seen as a "distributed system consisting of a collection of interconnected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources" SLAs are established and needed [11].
This results in the fact, that the SLA is seen as the foundation for the expected level of service between the consumer and the provider. Additionally, due to the dynamic nature of the cloud, continuous monitoring on QoS attributes is necessary to ensure SLAs [49] (see Figure 6.1).

In the following section of this report the main terms, such as CC, virtualisation, QoS and SLA are defined. Where as within CC it is shortly differentiated between the four

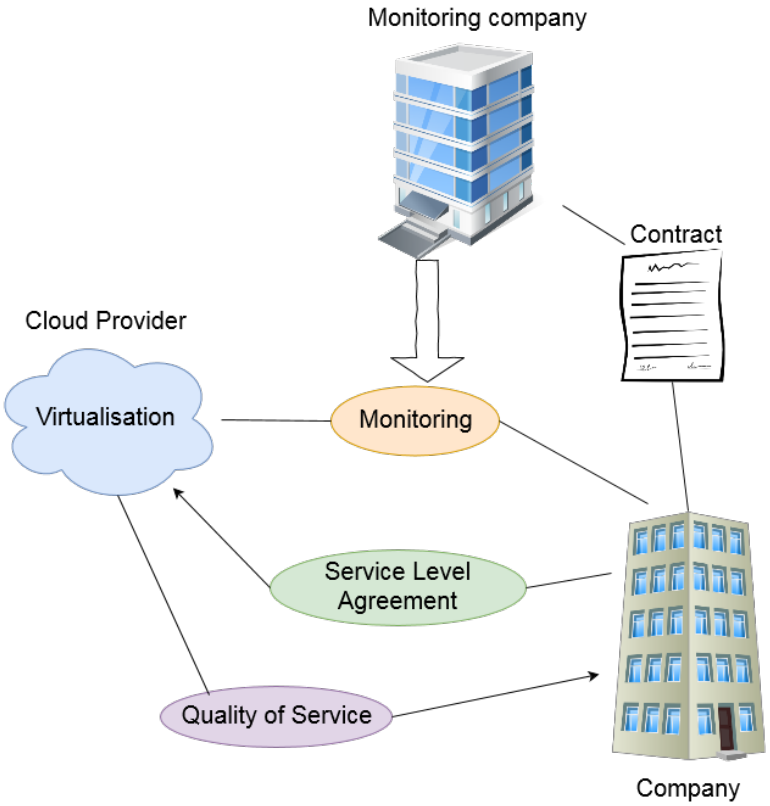


Figure 6.1: Participating entities in the problem

deployement models public, private, hybrid and community cloud. In the end of this section the SLA of Amazon EC2 is presented and compared to the SLA of Google Compute Engine.

6.2.1 Definitions

This section defines the main terms of this report such as CC including private, public, hybrid and community clouds, virtualisation, QoS and SLAs.

6.2.1.1 Cloud Computing (CC)

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics (On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured Service); three service models (Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS), Cloud Infrastructure as a Service (IaaS)); and, four deployment models (Private cloud, Community cloud, Public cloud, Hybrid cloud)." [39]

private cloud

"The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises." [39]

public cloud

"The cloud infrastructure is provisioned for open use by the general public. It

may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.“ [39]

hybrid cloud
“The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).“ [39]

community cloud
“The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises. “ [39]

6.2.1.2 Virtualisation

Virtualisation is defined as “a component within CC which allows separation of operating system from the hardware on which it is working“ [30]. Additionally, a virtual machine or more precise virtualisation helps to improve the efficiency of CC by providing the possibility to work on multiple operating systems and applications simultaneously [30]. All operating systems and applications are operating over the same physical server, which results in an increase of the utility and flexibility of hardware [30]. Not only the utility and flexibility of hardware within CC are improved by virtualisation, but also the energy saving [30].

6.2.1.3 Quality of Service (QoS)

“The collective effect of service performance, which determine the degree of satisfaction of a user of the service. The QoS is characterized by the combined aspects of service support performance, service operability performance, serveability performance, service security performance and other factors specific to each service.“ [27]
In other words, the QoS denotes the levels of performance, reliability, and availability offered by an application and by the platform or infrastructure that hosts it. Additionally, QoS is fundamental for cloud users, who expect providers to deliver the advertised quality characteristics. And for cloud providers, who need to find the right tradeoffs between QoS levels and operational costs. To measure QoS several related aspects of the network service are often considered, such as error rates, bit rate, throughput, transmission delay etc. [27, 43].

6.2.1.4 Service Level Agreement (SLA)

“An SLA is a contract between the provider of the service and the third party, such as purchaser of service“ [41]. In the contract the details of the service to be provided in terms of metrics agreed upon by all parties, and the penalties for meeting and violating the expectations are specified [11]. As a high grade of availability is needed for business purposes, the need to guarantee the availability and other metrics by contract rises continuously [8].
Described in other words, the role of an SLA is to clearly define service delivery expectations. Furthemore it provides an objective means of assessing if the performance meets service delivery expectations or not. This results in the fact, that it identifies the actions needed to improve the performance when it is requied. [8]

Example: SLA of Amazon EC2

In the SLA it is defined, that Amazon Web Services, Inc. try to provide a Monthly Uptime Percentage of 99.95% for Amazon Elastic Compute Cloud ("Amazon EC2") and Amazon Elastic Block Store ("Amazon EBS") (see Table 6.1). "In the event Amazon EC2 or Amazon EBS does not meet the service commitment (defined above), you will be elgibile to receive a service credit " [3, 2]. Thus, the Monthly Uptime Percentage is calculated by substracting from 100% the percentage of minutes during the month in which Amazon EC2 or Amazon EBS, is not available for the user. Not available means that more than one availability zone in which the user is running an instance, within the same regions, is unavailable to him or her. Unavailable and unavailability mean for Amazon EC2, when all of you running instances have no external connectivity and for Amazon EBS, when all of your attached volumes perform zero read write IO, with pending IO in the queue. Additionally, service credits are calculated as a percentage of the total charges paid by you for either Amazon EC2 or Amazon EBS for the monthly billing cycle in which the unavailability occurred (see Table 6.1). [3]

Comparison SLA of Amazon EC2 with Google Compute Engine SLA

Google Inc. define in their SLA that they will provide a Monthly Uptime Perecentage of at least 99.95% (see Table 6.2), which is exactly the same as in the SLA of Amazon EC2 (see Table 6.1). If Google Inc. does not meet this, they have a very similar procedure as defined in the SLA of Amazon EC2, but the term Monthly Uptime Perecentage is defined differently. It is defined as "total number of minutes in a month, minus the number of minutes of downtime suffered from all downtime periods in a month, divided by the total number of minutes in a month " [21]. Furthmore, the service percentage is defined as "percentage of monthly bill for the respective covered service which does not meet Service Level Objective that will be credited to future monthly bills of customer (see Table 6.2). " [21]

Table 6.1: Amazon EC2 SLA [3]

Monthly Uptime Percentage	Service Credit Percentage
Less than 99.95% but equal to or greater than 99.0%	10%
Less than 99.0%	30%

Table 6.1 shows the SLA of Amazon EC2 and Amazon EBS. The monthly uptime percentage is calculated by substracting from 100% the percentage of minutes during the month in which one of the services is unavailable for the user. The service credit percentage is a dollar credit, which is calculated as set forth below, that Amazon Web Services, Inc. may credit back to an eligible account if the SLA is breached [3].

Table 6.2: Google Compute Engine SLA [21]

Monthly Uptime Percentage	Service Credit Percentage
99.00% - < 99.95%	10%
95.00% - < 99.00%	25%
< 95.00%	50%

Table 6.2 shows the SLA of the Google Compute Engine. The monthly uptime percentage is calculated in the same way as in the SLA of Amazon EC2 and Amazon EBS. The service credit percentage is the percentage of the monthly bill for the respective covered service

which does not meet the SLA that will be credited to future monthly bills of the customer if the SLA is breached [21].

6.3 Problem Analysis

As mentioned in the introduction, the aim of this report is to give answers to the questions why the breaching of cloud SLAs happen and how to prove it. In this section it is given a short overview of the problem by discussing the involved entities.
The main participating entity in the given problem are CC providers. They offer their service through the following three different models.

Software as a Service (SaaS)
> "The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited userspecific application configuration settings." [39]

Platform as a Service (PaaS)
> "The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. " [39]

Infrastructure as a Service (IaaS)
> "The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls)." [39]

The second important entity is the client, which can be for example a private person or a company. In most cases, large companies use their own monitoring component such that they do not need to include a third party in the process. If a third party is needed a contract between the client and the monitoring company needs to be signed. Next, the connection between the cloud provider and the client is made by usage of an SLA, in doing so the cloud provider provides a specified QoS to the client. Concluding that the specific parameters between the provider and the client are monitored in order to detect agreement breaches after agreeing upon an SLA [34]. In the end, research wihtin the field of SLAs has pointed out many issues and challenges within SLAs, for example service cost related problems or that the specified amount of resources is not provided [41] (see Figure 6.1).
Problems are not only found between the relationships of entities but also within an entity itself. For example, the provider has problems such as network neutrality and outsourcing data. The provider wants to provide a good service, but still has to keep in mind the network neutrality, such that they cannot favor big companies before normal users.

In consideration of the given analysis of the participating entities, why are SLAs breached? Answered with the example of EC2, the performance on EC2 varies considerably and there exist several reasons why such performance inconsistencies may occur. The contention for non-virtualized resources (e.g. network bandwidth) is seen as the main reason for performance unpredictability in clouds. Additonally, another analysis has shown, that both small and large instances suffer from a large variance in performance. Furthermore, they have observed that one of the reasons of such variability is the different system types used by virtual nodes, e.g. Xeon-based systems have better performance than Opteron-based systems [51]. Finally, this question is discussed considering virtualisation and monitoring in the next section of this report.

6.4 Virtualisation: Problems and Solutions

The "most cloud service providers use machine virtualization techniques to provide flexible and cost-effective resource sharing among users" [58]. Virtualisation offers a lot of new opportunities, but has also some restrictions. There are two main goals using virtualisation technologies: saving money and get flexibility [18].
"There are two main types of virtualization technologies today a hypervisor-based technology including VMware, Microsoft Virtual Server, and Xen; and operating system (OS) level virtualization including OpenVZ, Linux VServer, and Solaris Zones" [47]. Most of the bigger cloud providers use the Virtual Machine Monitor (VMM) Xen [58].
Virtualisation of server systems brings an significant effort, when two ore more systems are merged, where the load peak is not at the same time [47] [14].
Xen is an open source paravirtualisation technology to create multiple virtual containers, so called domains, in one host system [51]. Xen is the most used virtualisation technology in CC nowadays. For example the Amazon EC2 is running on many Xen system [58]. Each domain runs an own instance of an operation system. It is also possible to run different operation systems on one host system [58]. For the access to the hardware, "the Xen hypervisor provides a thin software virtualization layer between the guest OS and the underlying hardware" [51].
"In general, hypervisor-based virtualization incurs higher performance overhead than OS-level virtualization does, with the benefit of providing better isolation between the containers" [51]. This means a hypervisor-based system is less scalable in increasing workload than a operating system level virtualisation, but it is much more flexible and the guest systems are more independend from each other [51].
In the next sections a few problems of virtualisation are announced and possible solutions for the problems named.

6.4.1 Memory

Sharing the main memory of the hosted systems can have a significant saving effect. With the running of multiple guest systems with the same operation system on the same host, a lot "of equal anonymous memory will be generated on the host hypervisor system" [7]. For this in Unix systems the Kernel Samepage Merging (KSM) is available. "The KSM main task is to find equal pages in the system" [7]. It adds the similar pages as shared and build to trees, a stable and an unstable, with the pages values. If there is a match between two equal pages, one of them can be released. If there should be an write event an a shared page, the page has to be dublicated first, so that the other user of the page has no influence in his data. Test resulted in a shared memory rate up to 750 MB out of the 2 GB for each of the two systems [7]. The usage of such technologies includes also dangers. For example when there are a lot of changing pages in one guest system, the

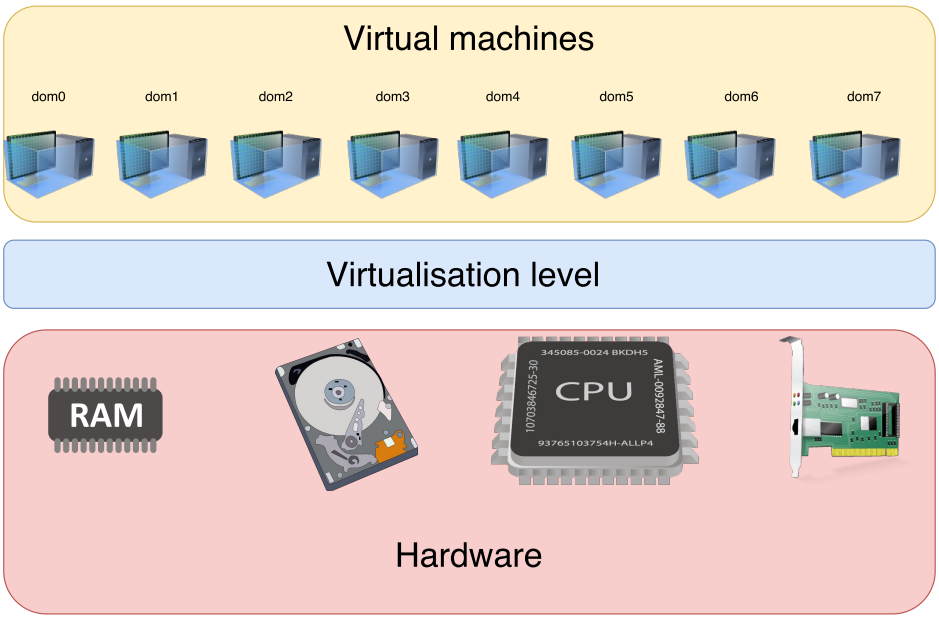


Figure 6.2: Virtualisation with a paravirtualisation technology

KSM has to duplicate them and therefore the system is endangered in running out of memory [7].
Not only the main memory can be the crux, also the L2 cache is well-known in Xen virtualisation system to cause a performance overhead [51].

6.4.2 Network Access

A research in Amazon EC2 showed a about 10 times longer answering time than non-virtualisied systems [58]. One problem is the Xen driver which supplies the network adapter to the guest system. "The communication between Xen driver domain and guest domain is done by copying data from memory to memory" [58]. So when multiple guest systems are delivering data to the network adapter, he has an overflow and data gets lost [58]. Another problem is the router safety check, which denies request up to 5 seconds before updating his tables [58]. But overall the Xen driver allows a throughput up to 5 Gb per second, if the network supports this [58].

6.5 Monitoring

Monitoring is the surveillance of a running system of compliance with given system parameters.
A cloud is a fragile system which has to be regulated and monitored in nearby real-time to get a constant QoS and therefore the comilance with the SLA [51]. For the end user the performance he gets is important, but the system consists of two parts: once the cloud farm of the cloud provider and the Internet or Network between the cloud and the user on the other side [49]. Therefore many different approaches exist to survey this aspects.
As an approach of standardizing the SLA agreements, the Web Service Level Agreement (WSLA) from IBM reseach provides an framework [38]. It defines four general scenarios used in almost every cloud application and has a point-of-view to the users sight. The first scenario is demanding for data by the server without further refinement, like providing a download stored on the server. The second one adds additional the request of a collection of data on the server. The third scenario requests customized and not only collected data, e.g using a database for gathering the informations. And in the fourth scenario the

customer additionaly specifies how the server has to collect the data. All this tests result in a measurement value of the response time and will be compared with a defined desired level [51].
In the next sections first the measuring of system parameters in the host systems are presented and afterwards the approch in monitoring the network connection. Some commercialy used systems to survey many aspects of a cloud system are presented.

6.5.1 System Measuring

System measuring in cloud systems is made in two different ways: measuring the host system or the virtual machine of the user in a IaaS or PaaS model. For the monitoring of the host system, the big CC sellers have developed their own systems [22].
A widely spreaded solution for monitoring private clouds or virtual machines (VM) is the open-source software Nagios licenced under the GNU GPL licence [33]. Nagios is a collection of modules and a core to initiate and log the tests. There are a bunch of modules available to survey Windows and Linux servers and internal network components like routers and switches. There are also components to survey websites, Domain Name System (DNS) configuration, HTTP and SSL certificates. Due to this possibilities it is suitable and used for monitoring web applications hosted in the cloud based on a SaaS or PaaS model [33]. In IaaS virtual machines with Nagios it is possible to monitor the SLA related to the CPU cores, the main memory availability and usage. There is also the opportunity to check the network bandwidth between different VM in the cloud [33].

6.5.2 Network Monitoring

Network monitoring is only with an online-system possible. This system is testing the connection between two nodes in the Internet in given time intervals. The most systems are monitoring the round-trip delay, the time the signal needs, and the loss of bursts [57].

6.5.3 Cloud Monitoring Systems

For monitoring cloud services and infrastructure a bunch of tools exists. These can be divided in 3 categories. The monitoring tools of the cloud provider, external tools giving an extern view combined with system parameters and monitoring tools performing request from outside in the way a user will use the service.
All big cloud providers are providing their own monitoring tools for their service. These tools have a deep insight of the hardware parameters of the host systems. They provide informations like the current CPU load, the memory usage, the internal network performance, ... and historical analysis of these. These tools are ideal to monitor a SLA reduced only to parameters of the availabiliy of the host system. It cannot monitor if the data center is reachable from outside. Another problem is the trust of the monitoring system because the party who is monitoring is the same as the party to be controlled.
Tools monitoring the sight of the user are widely spreaded. They are performing predefined actions which are similar to the user interaction. These tests are running on a platform independent from the cloud provider. Out of the results of the test it is only visible if a service is available and how long the response has taken. It is not possible to locate the problem to the cloud provider or the Internet between. An advantage of these tools is that all services the user needs are monitored included their supporting applications (e.g. databases).
The combinaison of these two approaches is also used. A monitoring service from outside the cloud provider is testing the system by sending periodic requests. Additionaly the system gets further information of the server by APIs. Often the data providers for these

APIs are insatalled on the guest systems and under control of the cloud user. The cloud provider often also provides an API for the values of the guest systems.
A lot of different systems from different suppliers exist. The following list gives a short, but not complete, overview of existing monitoring systems:

Amazon Cloud Watch [4]

AppDynamics End-User Monitoring [5]

AppNeta App View [6]

Bitnami Cloud Tools [9]

BMC Cloud Operations Management [10]

CA Unified Infrastructure Management [12]

Cisco Cloud Consumption as a Service [15]

CloudMonix [48]

DynaTrace Keynote Systems [17]

Exoprise CloudReady Monitor [19]

Google Stackdriver [24]

IBM SmartCloud Monitoring [25]

Idera Uptime Cloud monitor [26]

Kaseya Traverse [32]

Librato CloudWatch [36]

LogicMonitor [37]

Microsoft Azure Cloud Monitoring [40]

Monitis [42]

Nagios [44]

Netuitive [45]

Oracle Application Performance Monitoring Cloud Service [46]

RackSpace Public Cloud monitoring [55]

Riverbed Steelcentral Aternity [50]

ScienceLogic Hybrid IT Monitoring [53]

VMware Cloud Air Hybrid Cloud manager [56]

Zenoss Cloud Monitoring [60]

Out of these different solutions in the next sections 3 different solutions as an example for each type of tools will be presented in a more detailed way:

6.5.3.1 Cisco Cloud Consumption Service

Cisco Cloud Consumption Service is a Software-as-a-Service cloud monitoring tool. It is specifically designed for small and middle-size companies. This tool provides information about the actual state of the cloud, normally hosted by a public cloud provider. There are additional features compared to other monitoring platforms in recognising risks and duplicated systems. Cisco Cloud Consumption Service is specially designed for companies introducing their first monitoring system and using a lot of decentrialised cloud services, often in a SaaS manner [15].

6.5.3.2 Google Stackdriver Monitoring

Google Stackdriver is a new multi-feature tool from Google launched in 2016 for monitoring cloud services [23]. It is mainly designed to survey the Google Cloud Platform (GCP) and the Amazon Web Service (AWS). It has several features like a debugger and error reporting for the hosted applications, a monitoring tool, which can check multiple endpoints on their availabiliy, a tracing tool to survey the latency of the applications and multiple opportunities to alert in a case of dismatching with the desired values. For the usage also of non-technical users Google Stackdiver "provides a wide variety of metrics, dashboards, alerting, log management, reporting, and tracing capabilities" [22].
It's main advantage against other systems is the deep integration into the cloud platforms and therefore the wide-spreaded functions for an easy and fast monitoring of the hosted services. It is specifically designed for bigger firms hosting on both platforms GCP and AWS. There is a free version of Google Stackdriver and a commercial one. The free version has a heavily restricted scope of operation and supports only the Google web service [22]. Google Stackdriver was orignially developed for AWS and Openspace from Rackspace cloud services by the firm Stackdriver Inc. Google acquired the company in 2014 [35].

6.5.3.3 Kaseya Traverse

The solution suite Traverse from Kaseya is a multi-resource monitoring tool provided on a Software-as-a-Service (SaaS) base. It is specially "designed to monitor complex heterogeneous IT environments, including hybrid-cloud and virtual services" [32]. Hybrid-cloud means the usage of both, private and public cloud systems. Often companies use for default their own cloud in times of high load a system like Traverse can activate addiditional resources in a public cloud [32, 31].

6.6 Solutions and Discussion

First of all, the problem of breaching SLAs is only given for public and hybrid cloud, because there is no SLA for a private cloud. As introduced in the previous sections there are already a lot of monitoring tools available and used. A cloud system needs a continuous on-line monitoring to allocated the resources to the demand of the users [20]. With tools like Kaseya Traverse [32] operator of hybrid cloud systems can extend their own cloud by buying additional resources from a public cloud provider.
On the other hand, cloud providers "tempt to load as many accounts on the system as possible" [29]. So the customers are encouraged to monitor the cloud. In case of missfilling the SLA, a monetary compensation is often defined. To avoid spending this money, cloud provider have a incentive in providing enough resources. On the other hand, the rules of the market are playing for IaaS contracts. An virtualised system can be changed easier to another provider in case of many availability or resource problems with a cloud provider. Not only the cloud provider can influence the QoS for the end user, also the connection

over the internet between the cloud and the user has a significant influence [16]. Former studies "have shown that end-to-end Internet path performance degradation is correlated with routing dynamics" [57]. They researched the influence of failures in a system hosted in planetlab, a worldspreaded research network, in the border gateway protocol (BGP). The BGP "is the interdomain routing protocol that Autonomous Systems (AS) use to exchange information" [57]. AS are systems with a set of Internet Protocol (IP) networks. They were interested in the rates of packet loss, packet delay and the out-of-order packets due to routing failure, e.g. an overload on a link between two ISP, they recognised a rate of lost packets from up to 76 %. The delay, the time the packet needed from the sender to the receiver, was up to 19 seconds. 80% of all routing changes resulted in a loss of packets [57]. Schade, Dittrich and Quiane recognised in their survey that "contention for non-virtualized resources (e.g. network bandwidth) is clearly one of the main reasons for performance unpredictability in the cloud" [51]. Formulated the problem in an near-fashioned way: "The fundamental problem is that the simple textbook end-to-end delay model composed of network transmission delay, propagation delay, and router queuing delay is no longer sufficient. Our results show that in the virtualized data center, the delay caused by end host virtualization can be much larger than the other delay factors and cannot be overlooked" [57].
Optimizing the route on the Internet is difficult, because the network neutrality has to be guaranteed. This means, that internet service providers are not allowed to prefere signals from one customer against other ones [59]. A possible and efficient solutions for medium-sized and big companies is the renting of an own fibre-glass connection between the cloud data-center and their place of location. This would prevent routing problems and allow a stable and warranted connectivity.

6.7 Summary and Conclusion

The concept of CC has been analysed and discussed in this report in consideration of the following problem: why does breaching of cloud SLAs happen and how can we prove it? First of all it is given a short description of the participating entities in the given problem. Then the focus is set on the aspects of virtualisation and monitoring. Virtualisation is characterised by its flexibility and cost-effectiveness, additionally Xen is presented as one tool, which is used mostly by big providers. As the CPU, memory, RAM and network access are shared by virtualisation, it causes problems in limited resources for each VM. The aspect of monitoring is discussed in this report, because of its importance to regulate the cloud. It encourages the provider of the cloud to keep with the SLA, which is signed with the client. Measurements within the system are done by analysing the CPU and the memory load, concluding in measurements outside of the system by utilisation of ping, traceback, loss of packets and many other different tools.
Finally, it is concluded, that the breaching of cloud SLAs is not a big topic in literature and research. The reason is, because the market itself regulates the breaching of SLAs by monetary punishing non-agreements. Additionally, each customer can easily change the provider if they are not satisified with the available service, which is in most cases because of the routing over autonomous system to the customer and therefore its network neutrality.

6.8 List of Abbreviations

AWS Amazon Web Service

BGP Border Gateway Protocol

CC Cloud Computing

DNS Domain Name Service

EBS Amazon Elastic Block Store

EC2 Amazon Elastic Compute Cloud

GCP Google Cloud Platform

IaaS Infrastructure as a Service

IP Internet Protocol

ISP Internet Service Provider

KSM Kernel Samepage Merging

P2P Peer-to-Peer

PaaS Platform as a service

QoS Quality of Service

SaaS Software as a Service

SLA Service Level Agreement

VM Virtual Machine

VMM Virtual Machine Monitor

WSLA Web Service Level Agreement

Bibliography

[1] Alzahrani, A., Alalwan, N. & M. Sarrab: *Mobile Cloud Computing: Advantage, Disadvantage and Open Challenge*. In: Proceedings of the 7th Euro American Association Conference on Telematics and Information Systems (EATIS) 2014, Valparaiso, Chile.

[2] Amazon Web Services, Inc.: *Amazon EC2 Pricing*. December 2016. `https://aws.amazon.com/ec2/pricing/`.

[3] Amazon Web Services, Inc.: *Amazon EC2 Service Level Agreement*. December 2016. `https://aws.amazon.com/ec2/sla/?nc1=h_ls`.

[4] Amazon Web Services Inc.: *Amazon CloudWatch - Ueberwachungsservices fuer Cloud und Netzwerk*. December 2016. `https://aws.amazon.com/cloudwatch/`.

[5] Appdynamics, Inc.: *End User Monitoring*. December 2016. `https://www.appdynamics.com/product/end-user-monitoring/`.

[6] AppNeta, Inc.: *AppView Synthetic End User Web App Monitoring*. December 2016. `https://www.appneta.com/products/appview/`.

[7] Arcangali, A., Eidua, I. & C. Wright: *Increasing Memory Density by Using KSM*. In: Proceedings of the Linux Symposium 2009, Montreal, Quebec, Canada, pp.19-28.

[8] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I. & M. Zaharia: *Above the clouds: A berkeley view of cloud computing*. EECS Department, University of California, Berkeley. December 2016. `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html`.

[9] Bitnami, Inc.: *Bitnami Tools*. December 2016. `https://bitnami.com/tools`.

[10] BMC Software GmbH: *Cloud Operations Management*. December 2016. `http://www.bmcsoftware.ch/it-solutions/cloud-operations-management.html`.

[11] Buyya, R., Yeo, C.S. & S. Venugopal: *Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities*. In: Proceedings of the 10th IEEE Int. Conference on High Performance Computing and Communications, HPCC 2008, Dalian, China.

[12] CA Technologies, Inc.: *CA Unified Infrastucture Management*. December 2016. `http://www.ca.com/us/products/ca-unified-infrastructure-management.html`.

[13] Chen, Y., Li, X. & F. Chen: *Overview and Analysis of Cloud Computing Research and Application*. In: Proceedings of the International Conference on E-Business and E-Government (ICEE), 2011, pp.1-4.

[14] Chowdhury, N.M.K., & R. Boutaba: *Network virtualization: state of the art and research challenges*. In: IEEE Communications Magazine, 2009, Vol. 47(7), pp.20-26.

[15] Cisco Systems, Inc.: *Cloud Solutions - Cisco Cloud Consumption Service*. October 2016. http://www.cisco.com/c/en/us/solutions/cloud/cloud-consumption-service.html.

[16] Clark, D.D., Bauer, S., Lehr, W., Claffy, K.C., Dhamdhere, A.D., Huffaker, B. & M. Luckie: *Measurement and Analysis of Internet Interconnection and Congestion*. In: Proceedings of the Telecommunications Policy Research Conference (TPRC), 2014.

[17] Dynatrace, LLC.: *Mobile Apllication Monitoring*. December 2016. http://www.keynote.com/solutions/monitoring/mobile-app-monitoring.

[18] Etison, Y., Ben-Nun, T. & D.G. Feitelson: *A Global Scheduling Framework for Virtualisation Framework*. In: Proceedings of the IEEE International Symposium on Parallel Distributed Processing (IPDPS) 2009, Rome, Italy, pp.1-8.

[19] Exoprise Systems, Inc.: *Monitor your cloud and SaaS Applications*. December 2016. https://www.exoprise.com/solutions/monitor/.

[20] Galati, A., Djemame, K., Flechter, M., Jessop, M., Weeks, M. & J. McAvoy: *A WS-Agreement Based SLA Implementation for the CMAC Platform*. In: Economics of Grids, Clouds, Systems and Services: 11th International Conference, GECON 2014. Cardiff, UK, pp.159-171 http://dx.doi.org/10.1007/978-3-319-14609-6_11.

[21] Google, Inc.: *Google Compute Engine Service Level Agreement*. December 2016. https://cloud.google.com/compute/sla.

[22] Google, Inc.: *Stackdriver - Hybrid Monitoring*. December 2016. https://cloud.google.com/stackdriver/.

[23] Google, Inc.: *Introducing Google stackdriver: unified monitoring and logging for GCP and AWS*. December 2016. https://cloudplatform.googleblog.com/2016/03/Google-Stackdriver-integrated-monitoring-and-logging-for-hybrid-cloud.html.

[24] Google, Inc: *Stackdriver - Hybrid Monitoring | Google Cloud Platform*. December 2016. https://cloud.google.com/stackdriver/.

[25] IBM, Inc.: *IBM DevOps - Resources, case studies and best practices*. December 2016. https://www.ibm.com/cloud-computing/products/devops/.

[26] Idera, Inc.: *Uptime Cloud Monitor*. December 2016. https://www.idera.com/infrastructure-monitoring-as-a-service.

[27] International Telecommunication Union: *E.800: Terms and definitions related to quality of service and network performance including dependability*, August 1994. Updated September 2008 as Definitions of terms related to quality of service.

[28] Iosup, A., Ostermann, S., Yigitbasi, M.N., Prodan, R., Fahringer, T., & D. Epema: *Performance analysis of cloud computing services for many-tasks scientific computing*. In: IEEE Transactions on Parallel and Distributed systems, 2011, Vol. 22(6), pp.931-945.

[29] Jackson, K. R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J. (et al.) & N.J. Wright: *Performance analysis of high performance computing applications on the amazon web services cloud*. In: Proceedsings of the IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), 2010, Indianapolis, Indiana, US, pp.159-168.

[30] Jain, N. & S. Choudhary: *Overview of Virtualization in Cloud Computing*. In: Proceedings of the Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

[31] Kaseya Ltd.: *Traverse Developers Guide*. December 2015. `http://help.kaseya.com/webhelp/EN/tv/9010000/dev/EN_TraverseDevGuide_R91.pdf`.

[32] Kaseya Ltd.: *Unified Monitoring of Hybrid Cloud*. November 2016. `http://www.traverse-monitoring.com/`.

[33] Katsaros, G., Kuebert, R., & Gallizo, G.: *Building a service-oriented monitoring framework with rest and nagios*. In: Proceedings of the IEEE International Conference on Services Computing (SCC), 2011, pp. 426-431.

[34] Kosinski, J., Nawrocki, P., Radziszowski, D., Zielinski, K., Przybylski, G., & P. Wnek: *SLA Monitoring and Management Frameworf for Telecommunication Services*. In: Proceedings of the 4th International Conference on Networking and Services (ICNS), 2008, pp.170-175.

[35] Lardinois F.: *Introducing Google stackdriver: unified monitoring and logging for GCP and AWS*. December 2016. `https://techcrunch.com/2014/05/07/google-acquires-cloud-monitoring-service-stackdriver/l`.

[36] Liberato, Inc.: *Liberato + CloudWatch*. December 2016. `https://www.librato.com/cloudwatch`.

[37] LogicMonitor, Inc.: *LogicMonitor: SaaS-based Performance Monitoring Platform*. December 2016. `https://www.logicmonitor.com/`.

[38] Ludwig, H., Keller, A., Dan, A., King, R. & R. Frank: *Web service level agreement (WSLA) language specificatio*. IBM Corporation. 2003.

[39] Mell, P. & T. Grance: *The NIST definition of Cloud Computing: Recommendations of the National Institute of Standarts and Technology*. Gaitershburg, 2011.

[40] Microsoft, LLC.: *Microsoft Azure: Cloud-Computing-Plattform und -Dienste*. December 2016. `https://azure.microsoft.com`.

[41] Mirobi, G.J. & L. Arockiam: *Service Level Agreement in Cloud Computing: An Overview*. In: Proceedings of the International Conference on Control, Instrumentation, Communication and Computiational Technologies (ICCICCT), 2015.

[42] Monitis US, LLC: *Network & IT Systems Monitoring - Monitis*. December 2016. `https://monitis.com`.

[43] Motta, G., You, L., Sacco, D. & N. Sfondrini: *Cloud computing: the issue of service quality. An overview of cloud service level management architectures*. In: Proceedings of the 5th International Conference on Service Science and Innovation, 2015.

[44] Nagios Enterprises, LLC: *Nagios - Network, Server and Log Monitoring Software*. December 2016. `https://www.nagios.com/`.

[45] Netuitive, Inc.: *Homepage - Netuitive - Full Stack Performance Monitoring*. December 2016. http://www.netuitive.com/.

[46] Oracle, Inc.: *Application Performance Monitoring Service*. December 2016. https://cloud.oracle.com/application-performance-monitoring.

[47] Padala, P., Zhu, X., Wang, Z., Singhal, S., & K.G. Shin:. *Performance evaluation of virtualization technologies for server consolidation*. HP Labs Tec. Report, 2007.

[48] Paraleap Technologies, LLC.: *Cloudmonix | Advanced Cloud Monitoring & Automation Tools*. December 2016. http://cloudmonix.com/.

[49] Patel, P., Ranabahu, A.H., & A.P. Sheth: *Service Level Agreement in Cloud Computing*. Ohio, USA, 2009.

[50] Riverbed Technology Ltd.: *SteelCentral Aternity*. December 2016. https://www.riverbed.com/gb/products/steelcentral/end-user-experience-monitoring/steelcentral-aternity.html.

[51] Schad, J., Dittrich, J., & J.A. Quiane-Ruiz: *Runtime measurements in the cloud: observing, analyzing, and reducing variance*. In: Proceedings of the VLDB Endowment, 2010, Vol.3(1-2), pp.460-471.

[52] Schubert, L. & K. Jeffery: *Advances in Clouds*. In: Expert group report, Publications Office of the European Union, Luxembourg, 2012.

[53] ScienceLogic, Inc.: *Hybrid IT - Monitoring for All of Hybrid IT*. December 2016. https://www.sciencelogic.com/product/technologies/hybrid-it.

[54] Tawalbeh, L.A., Mehmood, R., Benkhlifa, E. & H. Song: *Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications*. In: IEEE Access, 2016, Vol. 4, pp.6171-6180.

[55] Rackspace US, Inc.: *Custom Infrastructure Monitoring*. December 2016. https://www.rackspace.com/de/cloud/monitoring.

[56] VMware, Inc.: *vCloud Air Hybrid Cloud Manager: VMware*. December 2016. https://www.vmware.com/ch/cloud-services/management/vcloud-air-hybrid-cloud-manager.html.

[57] Wang, F., Mao, Z. M., Wang, J., Gao, L., & R. Bush: *A measurement study on the impact of routing events on end-to-end Internet path performance*. In: ACM SIGCOMM Computer Communication Review, 2006, Vol.36(4),pp.375-386.

[58] Wang, G. & T.S. Eugene Ng: *The Impact of Virtualization on Network Performance of Amazon EC2 Data Center*. In: IEEE Infocom, Houston, USA, 2010.

[59] Wu, T.: *Network neutrality, broadband discrimination*. In: Journal of Telecommunications and high Technology law, 2003, Vol.2(141).

[60] Zenoss, Inc.: *Cloud Monitoring Software and Tools for Cloud-Based Networks*. December 2016. https://www.zenoss.com/product/what-we-monitor/cloud.

Chapter 7

Microtransactions with the Lightning Network

David Ackermann, Simon Bachmann, Philip Hofmann

Bitcoin is a decentralized payment system which allows anyone with an internet connection and a Bitcoin wallet to make transactions. However, Bitcoin implements transaction fees which render micropayments too expensive. Bitcoin has also insufficient speed and a slow confirmation time to become a widely adapted payment system. The Lightning Network is a proposal of Joseph Poon and Thaddeus Dryja to address the scalability problem of Bitcoin by using off-blockchain techniques. The purpose of this paper is to introduce and explain the idea behind the Lightning Network in simple terms.

Contents

7.1 Bitcoin

This section provides a motivation for Bitcoin, presents its current state, and discusses problems and potential solutions.

7.1.1 Motivation

At a very basic level the Bitcoin blockchain is a digital logfile that contains a list of all previously made transactions that happened in the Bitcoin network. People exchange money by updating this ledger. One of the core ideas of this system is to avoid any centralized control. Every node in the network maintains its own copy of the ledger, the digital logfile. Every validated transaction gets broadcasted over the network whereby every node updates his own ledger.
In today's environment, each bank maintains a centralized database for all of its customers. That database containins all crucial information of the customers relashionship to the bank, like their accounts, and their accounts balances. The fact that the ledger in the Bitcoin network is decentralized, instead of it being held by a single entity, invovles several important key differences.
First, in the traditional banking system, you only know about your own previous transactions, whereas in the decentralized ledger, you know about everyone's transactions within the network, although the system uses account numbers instead of the users names, which provides a low level of anonymity. There are other mechanisms which allow higher levels of anonymity in the Bitcoin network, but they will not be discussed in this paper.
Second, you trust your bank to maintain their ledger faithfully. It is a single responsible entity that can be sued if something goes wrong. This is not possible in a decentralized network, since there is no central, responsible entity in control.
Third, the order of events is more difficult to determine in a decentralized system. Broadcasting transactions over the network introduces some problematic drawbacks. Due to network latency, these broadcast messages might arrive in different orders. Two recipients might both think that their transaction was executed first and ship their product. This would allow the sender to spend these bitcoins twice - this double spending fraud is discussed in Section 7.1.2.4. Bitcoin decides on the transaction order by using the blockchain technology.

7.1.2 Current state

The current state is discussed within the following topics: Blockchain, miners, cryptography, and double spending.

7.1.2.1 Blockchain

When a new transaction is created, it enters a pool of unconfirmed transactions. Each miner competes for the privilege to include those unconfirmed transactions in the blockchain. To do this, each miner needs to solve a encryption challenge. This encryption can only be solved by brute forcing (random guessing) the input.
When this input is guessed correctly, all the transactions need to be verified. To verify a transaction, the miner checks all the previously made transactions. By adding up all the unspent bitcoins, one can verify if the sender is in possession of the amount that he wants to send. The miner then includes those verified transactions in a block and the block then enters the blockchain and is linked to the previous transaction block. This results in a chain of transaction blocks that are linked together - the blockchain. Since each block references to its previous block, the system can distinguish the order of transactions in time. Transactions in the same block are considered to have happened at the same time.

7.1.2.2 Miners

Miners are the nodes in the network that dedicate their computing power to verify transactions and making sure that everything that is happening in the Bitcoin network is legitimate. As mentioned in Section 7.1.2.1, a block of transactions must be confirmed before it can be added to the blockchain.
In the Bitcoin network, a single entity that determines the validity of all transactions is not wanted. Instead random miners from the entire network are competing to confirm a block and add it to the blockchain. The encryption challenge guarantees that a miner is chosen randomly. The miner competes by computing the correct input for the hashed block. The more miners are in the system, the smaller is the probability that the same node guesses correctly multiple times in a row. Therefore, the system becomes more secure with a larger user base.
The node which solves the hash has the privilege of adding the block onto the blockchain and receives a monetary reward. It would take several years for a typical computer to solve the hash. The chances of solving the hash and receiving the reward - which takes in average 10 minutes for the entire network - is extremely low for an individual with a large user base.
That's where mining pools come into place. A mining pool is a group of miners that provide their mining equipment to guess the hash with non-competing computer power to have better chances for guessing the input correctly. The reward is then split among the miners in the pool [3].

7.1.2.3 Cryptography

Bitcoin requires a signature to prove that the sender is the real owner of that account. It uses the mathematical concept of public and private keys. A private key allows the user to create signatures and the connected public key allows others to verify this signature without actually knowing the private key. In Bitcoin, public keys are used as the address of a transaction. If you want to send money to Alice, you send it to her public key.
To spend money from a public key address, the sender must verify that he is the true owner of that address without revealing his own private key. He creates a digital signature with his private key and the transaction message with a hash function. This signature can be checked by other people in the network using a different function. They now can verify the correspondence between the digital signature and the private key without having access to the private key. By including the transaction message when creating the digital signature, the signature will be different for every transaction and cannot be reused by someone else. Further, the transaction message cannot be modified by someone in the network because this would result in an invalid signature.

7.1.2.4 Double Spending Problem

One of the main difficulties of a distributed ledger is the double spending problem. Assuming Alice has exactly 1BTC in her Bitcoin wallet. She wants to trick Bob on sending her a new laptop without Bob ever receiving any money. To achieve this, she creates two transactions - one transaction with 1 BTC to Bob and 1BTC to Carol (some third party). Both transactions are added to the pool of unconfirmed transactions. Bob's transaction is added to a block and the hash attached to this block is solved. Bob sees that the transaction is added to the blockchain and sends the laptop to Alice.
There is the possibility that a block is mined before that miner received an updated version with Bob's transaction of the blockchain. At that point in time there exist multiple blockchain branches. If this is the case, only the largest branch is considered as the valid

branch and all the other transactions from the shorter branch would be added to the pool of unconfirmed transactions.
Therefore, if Alice could compute a longer branch than the one where the transaction to Bob remains, the transaction to Bob would be thrown back into the pool of unconfirmed transactions. It would no longer be valid since Alice spent all her money to Carol. Bob does not get any money and the laptop is already on its way to Alice.
If Alice wants to present a longer branch just after Bob has sent the laptop, she races the entire network and needs an enormous amount of computing power. The maximum number of consecutive blocks added to the blockchain was done by a large mining pool with six blocks. Therefore, if Bob waits more than 60 minutes before sending the laptop it is nearly impossible for Alice to defraud Bob.

7.1.3 Problems

- **Scalability**: Bitcoin is not scalable. Bitcoin can currently only process a theoretical maximum of 7 transactions per second (tps). But in real world conditions the maximum lies between 2 to 3 tps. In comparison to *Paypal* [9], they process almost 11.5 million transactions every single day. This results in a processing power of 133 tps. Also the *Visa* [10] network processes an average of 150 million transactions every day. Which leads us to an avarage of 1736 tps. This illustrates that Bitcoin must find a solution that is more scalable to serve a similarly sized customer base.

- **Transaction costs**: Bitcoin is not suited for micro transactions. The transaction fees are too high. According to *New Service Finds Optimum Bitcoin Transaction Fee* [2] the average size of a Bitcoin transaction is 645 bytes. The currently most popular fee ratio is CoinTape, they charge 41—50 satoshis per byte. This results in a 0.16 USD fee for every single transaction, which is too high for a microtransaction.

- **Confirmation time**: Confirmation time of a transaction is too long. It can vary greatly between 10—60 minutes. This is not suited for places where instant payments are required.

7.1.4 Possible Solutions

To tackle the transaction cost issue, the following three solutions are presented: centrarized server, sidechains, or payment channels.

7.1.4.1 SQL Database Model

Whoever is in possession of the private key for a bitcoin address can manage this address. If you run your own wallet application, you are the only one who has access to the private key. You are also fully responsible to keep it that way.
There are online services such as Coinbase or Circle that work similar to a bank. They worry about security, own all your bitcoins and your account balance represents a promise that they owe you the number of bitcoins that is stored in their database.
This solves the scalability issue of the Bitcoin network to some extent, because transferring money from a Coinbase account to another Coinbase account occurs off the blockchain. This allows scalable, cheap and fast transaction from Coinbase user to Coinbase user.
One of the core motivation of Bitcoin is that you avoid a centralized database. With such a system like Coinbase the decentralized idea of Bitcoin and no trust policy is gone to some degree. You need to trust Coinbase that they act faithfully.

7.1.4.2 Sidechains

Bitcoin sidechains exist alongside the main Bitcoin blockchain. They allow bitcoins and other assets to be transferred between blockchains. If two parties on the same sidechain make a transaction, it is not recorded on the main blockchain and therefore, the traffic on the main blockchain can be relieved. Although, this is not primary a solution to the scalability issue of Bitcoin. Sending funds between two blockchains results in two transaction if these blockchains are not linked directly to each other.

7.1.4.3 Payment Channels

Since the opening and closing transaction of a payment channel need to be recorded on the blockchain, payment channels only solve the scalability issue if two parties exchange funds regularly. Payment channels using smart contracts guarantee fast and cheap off-blockchain transactions without the need of trust.
The Lightning Network [1] is an extended solution to a traditional payment channel. It uses open channels between nodes in the network to established multi-party payment-channels. How this can be done without the need of trust and without being recorded on the blockchain is discussed in the following Section.

7.2 Lightning Network

The Lightning Network addresses major flaws of Bitcoin:

- **Scalability**: Since Bitcoin is not scalable due to its low transaction throughput, the Lightning Network uses payment channels to have less transactions on the blockchain. It uses smart contracts to guarantee the no trust policy. This allows the network to handle millions to billions of transactions per second.
- **Transaction costs**: With less transactions on the blockchain the transaction costs become a fraction of the original costs.
- **Confirmation time**: The confirmation time of a transaction within a payment channel is nearly instant. As soon as a payment channel between to parties is opened they can exchange money within milliseconds.

7.2.1 Payment Channels

There are several building blocks that need to be understood before diving into the Lightning Network. These building blocks are necessary to understand the fundamental element of the Lightning Network: a bidirectional payment channel. Detailed information is provided by [4, 5, 6].

7.2.1.1 Multisignature Address

Multisignature addresses - or simply multisig - are shared Bitcoin addresses that require multiple signatures to spend bitcoins from. The Lightning Network mainly uses 2-of-2 multisigs. A 2-of-2 multisig is a shared Bitcoin address between two Bitcoin users where both must authorize (sign) every transaction. In general, any M-of-N multisig address up to 15 participants would be possible in the Bitcoin Network.
In Figure 7.1 Alice and Bob previously created a multisig address to which both hold a key. Alice wants to spend 5 bitcoins from this multisig back to her own address. Therefore, she creates and signs a transaction and passes it to Bob. If Bob agrees to this transaction

Figure 7.1: Multisignature address

the transaction is valid and 5 bitcoins can be spent. If Bob does not agree Alice will not be able to execute this payment. Red color is used for Bob, blue color for Alice (and green color for Carol) throughout the entire paper.

7.2.1.2 Secrets and Hashes

A secret is a randomly generated string. As the name suggests, a secret should ideally not be easy to guess. Accordingly, it has to be large and complex. A cryptographic hash function is a mathematical one-way algorithm, that takes a string (in this case the secret) as an input and computes it into a hash, a unique string of numbers. The essential property of this concept is, that someone who knows the secret can easily recreate the hash, but the secret cannot be reproduced from the hash. In Figure 7.2 a transaction is illustrated that is using secrets and hashes. Colored hashes and secrets indicate to what person the hash or secret belongs to.

7.2.1.3 Time Locks

A time lock delays the execution of a transaction from a specific address. More precisely: the signature of a user on this address for a specific transaction becomes valid only after that time lock expires. There are two types of time locks: an absolute and a relative type.

- The relative type is called a CheckSequenceVerify (CSV), it refers to an event, rather than a specific point in time. That event would be a specific number of blocks added onto the blockchain.
- The absolute type is called a CheckLockTimeVerify (CLTV), it refers to an actual time and date. With this type of lock one can create a hash time locked contract (HTLC) as in Figure 7.3 illustrated. Alice makes a transaction to a new multisig address. Alice wants to be sure that Bob can only unlock this multisig if he provides the secret from another transaction (green key). She also wants to make sure that

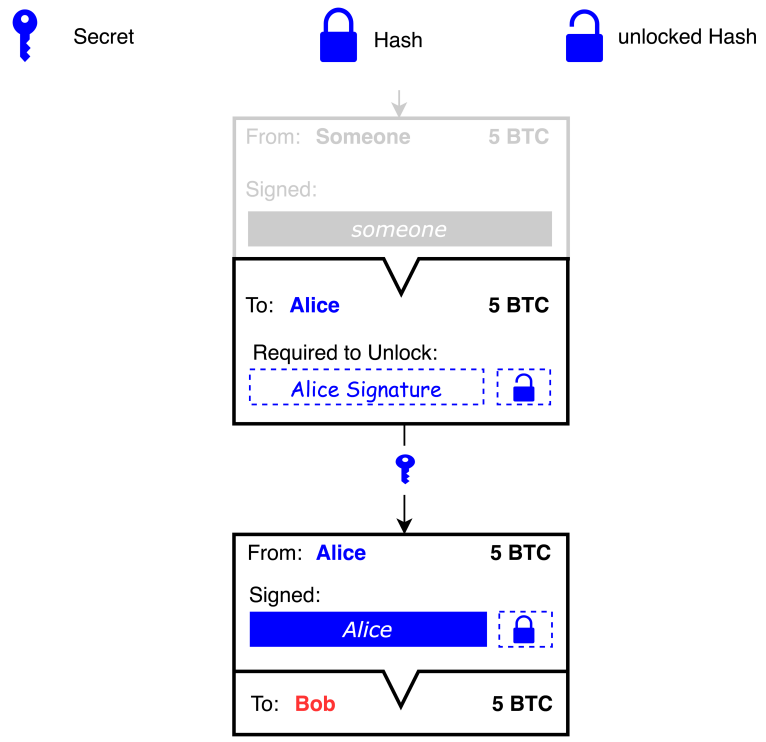


Figure 7.2: Hashes and values

she gets her bitcoin back if Bob does not corporate. That is why the CLTV lock is needed.

7.2.1.4 Bidirectional Channel

The idea of payment channels has been around for a while, but only with limited use. They are one-directional. Alice can pay Bob through several off-chain transactions, but Bob can't pay Alice at all. To solve this problem, the concept of a bi-directional channels was created. This bi-directional channel is realized with a 2-of-2 multisig address.

Opening the Channel

Consider the following scenario: Alice wants to send 1 BTC to Bob. Since they both expect to exchange funds regularly, they decide to open a payment channel.

To setup a bi-directional channel, Alice and Bob have to both first agree on an opening transaction, which specifies how many bitcoins each deposits into the channel/multisig address. In this example they both deposit 5 bitcoins into a 2-of-2 multisig address. This is the **opening transaction**.

Now Alice and Bob both generate a secret and the corresponding hash. Alice now creates a **commitment transaction**. In this transaction, Alice would send 4 bitcoins to herself, and 6 bitcoins to a separate 2-of-2 multisig address. Alice signs this transaction and gives it to Bob. Bob does the same as Alice but mirrored: He also creates a commitment transaction, where he sends 6 bitcoins to himself and 4 bitcoins to a separate multisig address. He signs this transaction and gives it to Alice.

These commitment transactions serve similarly to a contract: They are a trustless assurance that the owner possesses his/her fair share of the multisig address.

Alice and Bob both sign their opening transactions and broadcast it onto the blockchain. The channel is now open. At this point, both have the option to sign and broadcast the commitment transaction and thereby close the channel. But the whole trick is that neither of them signs and broadcasts the commitment transaction. The channel stays open

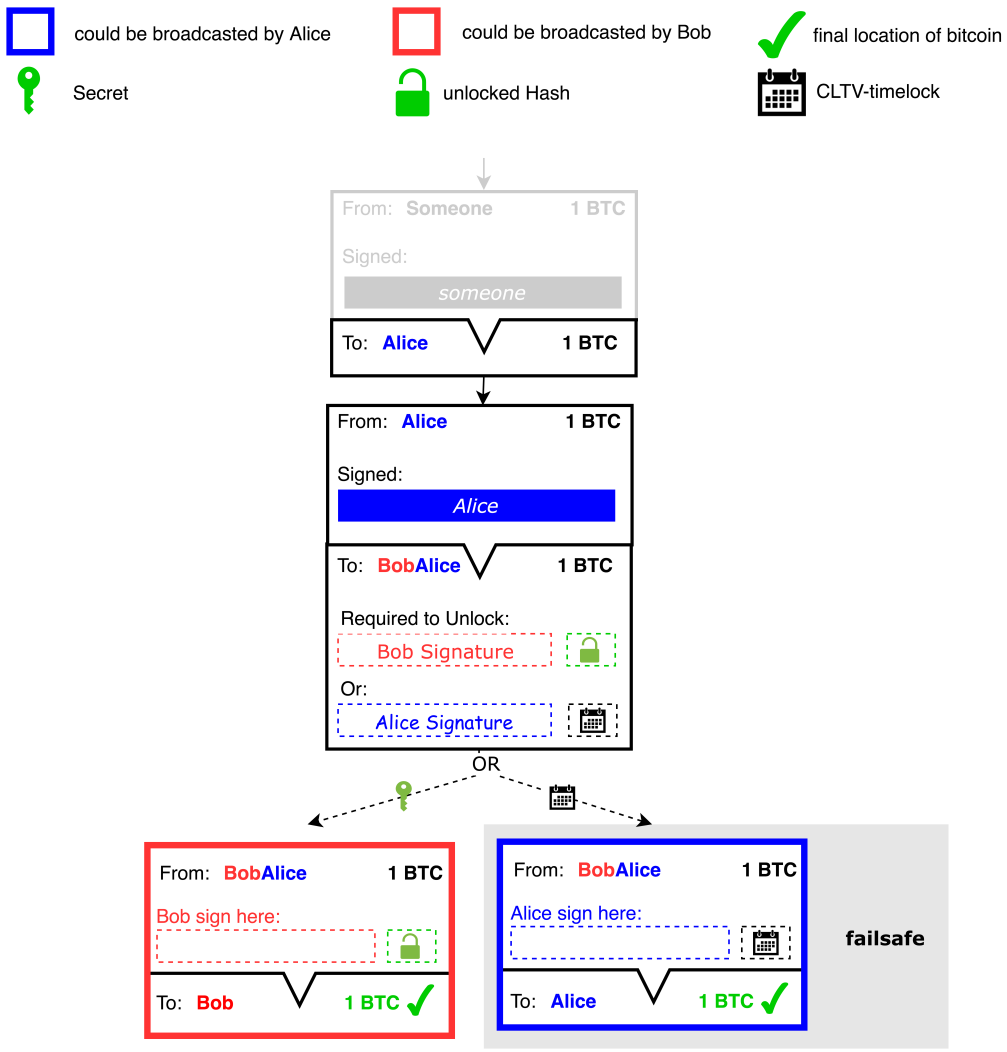


Figure 7.3: Hash Time Locked Contract

and can be used for further transactions between them. In Figure 7.4 the just described scenario is illustrated.

Updating the Channel

A bi-directional payment channel is now established. In the next scenario, Bob wants to pay Alice 0.5 bitcoins back. To do this, they both exchange the secrets from the previous transaction which effectively invalidates it.
Afterwards, they generate a new secret and again exchange half-signed commitment transactions: Bob creates a commitment transaction where he immediately gets 5.5 bitcoins and 4.5 bitcoins go into a 2-of-2 multisig (same setup as before). Alice does the same but mirrored. For every new transaction, this process will be repeated. The updated transaction is shown in Figure 7.5. This mechanism is called a **failsafe** (see Section 7.2.1.4).

Closing the Channel

If both parties want to close a channel by mutual consent they can both sign a new transaction, that gives them their fair share of the multisig. Everybody will get their amount of bitcoins immediately after it has been added to the blockchain. Every previous transaction in the channel can be discarded.
The channel will also be closed in case either Alice or Bob decides to broadcast their commitment transaction.

One can see in the examples above, Alice or Bob needs to unlock the half-signed transaction they received in order to broadcast a transaction. So Bob could sign the transaction, that he has received and that has already been signed from Alice. He then broadcasts this transaction. Alice receives her bitcoins immediately (will be included on the blockchain as soon as possible). Bob needs to wait for the required time lock, then his transaction will also be added to the blockchain. The channel is now closed.
If Bob broadcasts an old transaction, Alice could with Bob's private key to this transaction also get Bob's money. In this scenario, there is no time lock restriction. Alice receives the all the money from the multisig right away and the channel is closed.

Fraud Prevention
Bob had 6 bitcoins in the first transaction and only 5.5 bitcoins in the second. What prevents him from not cooperating and insisting on the first transaction?
If Bob decides to sign the first commitment transaction Alice gave him, Alice gets 4 bitcoins immediately. The second transaction is from the multisig to Bob's account - but there's a time lock of 1000 blocks on it. Since Alice's transaction got included in the blockchain and she received her 4 bitcoins, she knows that Bob tried to broadcast an old transaction. Bob's transaction is delayed and Alice can now use Bob's previously exchanged secret to receive his funds immediately. Alice has now all the money from the channel and the channel is closed. This mechanism strongly encourages cooperation and has no trust in it.

7.2.1.5 Multiple Party Channel

Alice and Bob already have an open channel and Bob and Carol have an open channel. What happens if Alice wants to send Carol some bitcoins? Using smart contracts, it is possible to send bitcoins from Alice to Carol via Bob, without needing trust.
The following scenario is visualized in Figure 7.6. To initiate the payment, Alice tells Carol to create a secret. Carol sends the hash to Alice and Alice (1) forwards it to Bob (2). Now Bob pays Carol 1 bitcoin and gets the secret to the hash in exchange (3). Seeing Bob having the secret, Alice can be assured that Bob paid 1 bitcoin to Carol. Alice now pays Bob 1 bitcoin and gets the secret in exchange (4). The payment is now successfully completed.
The problem with this approach is, that there's still trust involved: How can Bob be sure that Carol gives him the secret after sending 1 bitcoin?
This is where CTLV-time locks come in: Instead of just giving 1 bitcoin back, the bitcoin is deposited into a 2-of-2 multisig which can be unlocked either with the secret or after a certain period of time.
Carol provided Alice with the corresponding hash to her new secret. Since Alice will pay Carol through already existing channels, she will use her channel with Bob as an intermediate. Therefore, she provides Bob with Carol's hash, updating their payment channel with a new commitment. But in this new commitment, Alice will separate a bitcoin, which is her payment to Carol. This bitcoin is locked up with Carol's hash. Bob will only be able to get this bitcoin with Carol's corresponding secret. To get this secret from Carol, he sends Carol an update request of their payment channel. Bob now creates a commitment on their payment channel with Carol, where he also locks up a bitcoin, Alice's payment for Carol. Carol can unlock this with her corresponding secret. So she sends her secret to Bob to get her bitcoin. Bob forwards this secret to Alice to get his bitcoin.
Since Carol sent Alice the hash, she is actively listening for an update request from any source. Bob is acting as a middle man since he received Carol's hash from Alice. Therefore, he sends Carol an update request. In this update, Bob separates or locks up a bitcoin,

which represents Alice's debt to Carol. This separated bitcoin can be unlocked if Carol provides the correct secret to Bob. In Figure 7.7 a complete overview with all possible failsafes is shown.

7.2.1.6 Fraud Prevention

The mechanism to prevent non−cooperation is very similar to before with direct bi-directional channels. Additionally to give Bob additional security the time lock for the multisig with Carol will always run out before the time lock of his multisig with Alice. This prevents the situation where he pays Carol 1 bitcoin and getting the secret but being unable to get the bitcoin back from Alice.

7.2.1.7 Privacy Issues

The routing nature of the Lightning Network provides routing nodes with more data than others, giving them an edge over competitors. This is similar to an ISP which accumulates data of its customers. It has exclusive data in comparison with the other ISPs and can use this data for intern statistics, marketing purposes or sell it to other companies. The routing nodes in the Lightning Network also accumulate data of the channels and the transactions that flow through this channel. Since it is not public data (like in Bitcoin), it gives them an edge over competitors and it is unknown of what they are using this data for.

7.2.2 Scalability

As previously discussed, the Bitcoin network has scalability issues. It is therefore not suited to be used by the masses as for example Paypal or Visa is used today. The Lightning Network tries to overcome those obstacles. The Lightning Network takes a lot of transactions off the blockchain. The transactions will be stored locally in the nodes. The problem with this whole scenario is the processing power that is needed to include a lot of transactions into the blockchain. There are currently between 2 to 3 transactions per second processed by the bitcoin miner network whereas in comparison the visa network handles an average of 1736 transactions per second (see Section 7.1.3). The Lightning Network only uses the blockchain for the following use-cases: Opening a channel through a broadcasted opening transaction, the closing of the channel through a broadcast of any transaction from that channel.

7.2.3 Costs

Bitcoin has a minimal transaction fee defined since 2012 of 0.0005 BTC [11]. If there were transactions with no transaction fees, two substantial problems occur. First, one could be able to spam the bitcoin network with feeless transactions. Second, one could send a feeless transaction into the network, but every miner would reject to validate this transaction. Therefore, it would never make it into the blockchain.
In a perfect world, payment channels do not have to be closed at all, and the transaction fees in the Lightning Network withing open channels are effectively zero. This makes the Lightning Network attractive. Since almost no transaction has to be broadcasted, almost no costs are generated. If a channel has to be closed for some reason, only the newest commitment transactions have to be considered - all the previous transactions can be discarded. On a cost per transaction basis, the Lightning Network is much cheaper than broadcasting every single transaction.

7.2.4 Use Cases

The following list of use cases shows the potential of the Lightning Network.

7.2.4.1 Instant Transactions

In contrast to the Bitcoin network, the Lightning Network requires no confirmation of transactions.Therefore, transactions within the Lightning Network can be considered instantaneous, they are valid as soon as the channel is updated with another transaction.
One can pay for a cup of hot chocolate at Starbucks instantaneously, without people in the long line getting mad at confirmation waiting. One can buy the ticket for the train ride with the train already approaching. One can pay the parking ticket before even getting into the car.

7.2.4.2 Content Seeding

While popular torrents have no problem finding seeders, more rare content is often hard to find. Adding a financial component to seed content, the access to rare content would increase quite drastically.

7.2.4.3 Ads

The hassle for content creators to get paid online would decrease dramatically. Publishers could get paid instantly for showing ads on their websites. Youtube could pay content creators in real-time according to views. Or one could bypass ads for a small amount, which is paid instantly. You could also get paid to watch ads if you're willing to watch them.

7.2.5 Possible Risks

With all the advantages of the Lightning Network, there are risks as well. In the following, these are presented.

7.2.5.1 Centralization

There is a discussion about centralization risks that lead to potential super-nodes. There are several aspects that speak for and against a potential centralization. On one hand, one can make money when one route payments from one node to another. Also, one has some interesting data at hand, that could potentially be analyzed and sold to marketing firms. These are both incentives for building well-connected nodes.
On the other hand, a supernode is more vulnerable to attacks, your money is locked up in channels, if other parties are not cooperating and the revenue of routing payments is marginal and will fade away as soon as the network established itself.

7.2.5.2 Scaling Transactions, not Users

Although the Lightning Network seems to be one of the best solutions to scale the Bitcoin Network, it only prevents transactions from entering the Blockchain and does nothing directly to scale to more users. The Lightning Network still requires on−chain transactions to open and close channels, which is limited by the maximum block size of Bitcoin. The Network will certainly help scale Bitcoin, but it would just delay the maximum block size discussion, which has to be solved eventually.

7.2.5.3 Failure Mode

When there are a big number of contracts that need to be settled at the same time there's a potential failure. There's only limited amount of data that can go through the blockchain. If there's a large number of channels, that close out rapidly you can run out of capacity. This means the cost of doing a transaction could rise substantially until people start losing out. The commitment transaction and its time lock is a fundamental part of the lightning network. If one party tries to cheat and broadcasts an old commitment transaction the other party needs to access the mutual multisig and broadcast the transaction before the time lock expires. If this is not possible the whole construct fails.

7.2.5.4 Channels expose you to market volatility

If Bitcoin were the dominant currency in the country, having bitcoins locked up in a channel would not be a big problem. You could just hold on to the currency and wait for the rise in value, due to the deflationary nature of Bitcoin. If the market experiences a sharp correction a large amount of people want to close out their channels. This hits the blockchain all at the same time and leaves people waiting for their confirmation, due to the limited amount of traffic. This is a known problem with Bitcoin. But having your money locked in channels means selling your bitcoins in favor of another currency takes more time, so your exposure to market volatility increases.

7.2.5.5 Data Loss

Through data loss of one party, it is possible for the party to steal the funds of the channel. The party with the data loss does not know the current state of the channel, but also not the stack of old commitments and the keys for the failsafe's. The party with the data can now dig out an old commitment that is most favorable and broadcast it onto the blockchain.

7.2.5.6 Unnoticed Broadcast

In order to use the failsafe method on a counterparty that broadcasts an old commitment, one has first to notice this occurrence. If the time lock of the counterparty has run out, there is no way for you to get your money back. A possible solution would involve a third party together with a monetary incentive to watch the channel and inform either party in case of a broadcast of an old commitment.

7.3 Ethereum

Ethereum [8] uses blockchain technology to represent the ownership of property, even though it is very different to Bitcoin. Bitcoin is mainly known as a crypto currency. Ethereum is more than a currency, it is a decentralized platform that runs smart contracts, a blockchain app platform in other words. It is a more general approach than another cryptocurrency. Developers can create markets on this network, store registries of debt, store future contracts and many more things.
The Ethereum Wallet is the gateway, it allows you to store ethers (a crypto-asset built on Ethereum) but also to write, deploy and use smart contracts. It is also different to Bitcoin in the sense that Ethereum uses accounts with balances as the fundamental object (in contrast to chained transactions balances).
Ethereum even has its own programming language called Solidity. It is similar a high-level language, designed to target the Ethereum Virtual Machine, and similar to JavaScript in its syntax.

7.3.1 Ether

Ether is a crypto-asset built on Ethereum. It describes itself as a crypto-fuel, keeping the whole Ehtereum network running. It is a token, whose purpose is to accommodate for the resources; a payment from the clients to the machine resources to execute their commands. It poses itself as an incentive for developers to write high-quality, clean code and for resource contributors to keep the network running.

7.3.2 Possibilities

Ethereum with Solidity enables the following: One can design and issue its own cryptocurrency. One can create digital tradeable tokens. This digital token can represent any tradable good like gold certificates, in game items, coins, loyalty points, virtual shares, etc.
Another possibility is to kickstart a project with a trustless crowdsale. One can also use this network for a crowdsale to sell virtual shares in a blockchain organization or to auction a limited number of items.
Another possibility is to create a democratic autonomous organization with the Ethereum Network. This is a virtual organization run by a robot, where members vote on issues, a democracy on the blockchain. The way this works is that this organization has an owner, a president so to say. The owner manages the voting members (adding and removing). Any member makes a proposal by a transaction or by executing some contract, and the other members can vote in favor or against the proposal. After a predetermined number of members has voted in favor, the proposal can be executed.

7.3.3 Raiden Network

The Raiden Network [12] is similar to the proposed Lightning Network of Bitcoin as it leverages off-chain state networks. Its aim is to increase scalability (1'000'000+ transfers per second possible) and establishing low transaction fees, allowing for micropayments to be effectively used. A Raiden node runs parallel to an Ethereum node. It networks with other Raiden nodes to facilitate transfers. It also exchanges data with the Ethereum blockchain to manage deposits. The Raiden Network in this sense lives within the Ethereum Network.
There are some differences between the Raiden Network and the proposed Lightning Network. First, the implementation of this concept on the Ethereum Platform is easier compared to Bitcoin, because of the Ethereum Virtual Machines. Another difference is that the channel funds are fixed in the Lightning Network and to add funds to such a channel, the channel needs to be closed and another one opened with the desired funds. In the Raiden Network, funds can be added to already existing channels. If Alice wants to increase funds in the channel, she can send funds directly to her side of the channel, increasing thereby the total funds in the channel [7]. The payment channels within Ethereum allow for balances in one direction to go negative, if both parties agree. This would be considered more of a credit channel than a deposit channel.

7.4 Conclusion

The Lightning Network is a proposed solution to make Bitcoin more scalable. Even tough it is well thought out, it is still a concept that needs to be deployed. The risks and problems that will be introduced by the actual deployment cannot be foreseen, and therefore, its effect in the real world are unclear. Concept wise, it can be concluded that it has its merits, but does not provide a silver bullet to the bitcoin scalability problem.

It tries to take most transactions off blockchain, but is therefore not addressing Bitcoins underlaying problem, which resides in the blockchain itself.

Figure 7.4: Opening and Commitment Transaction

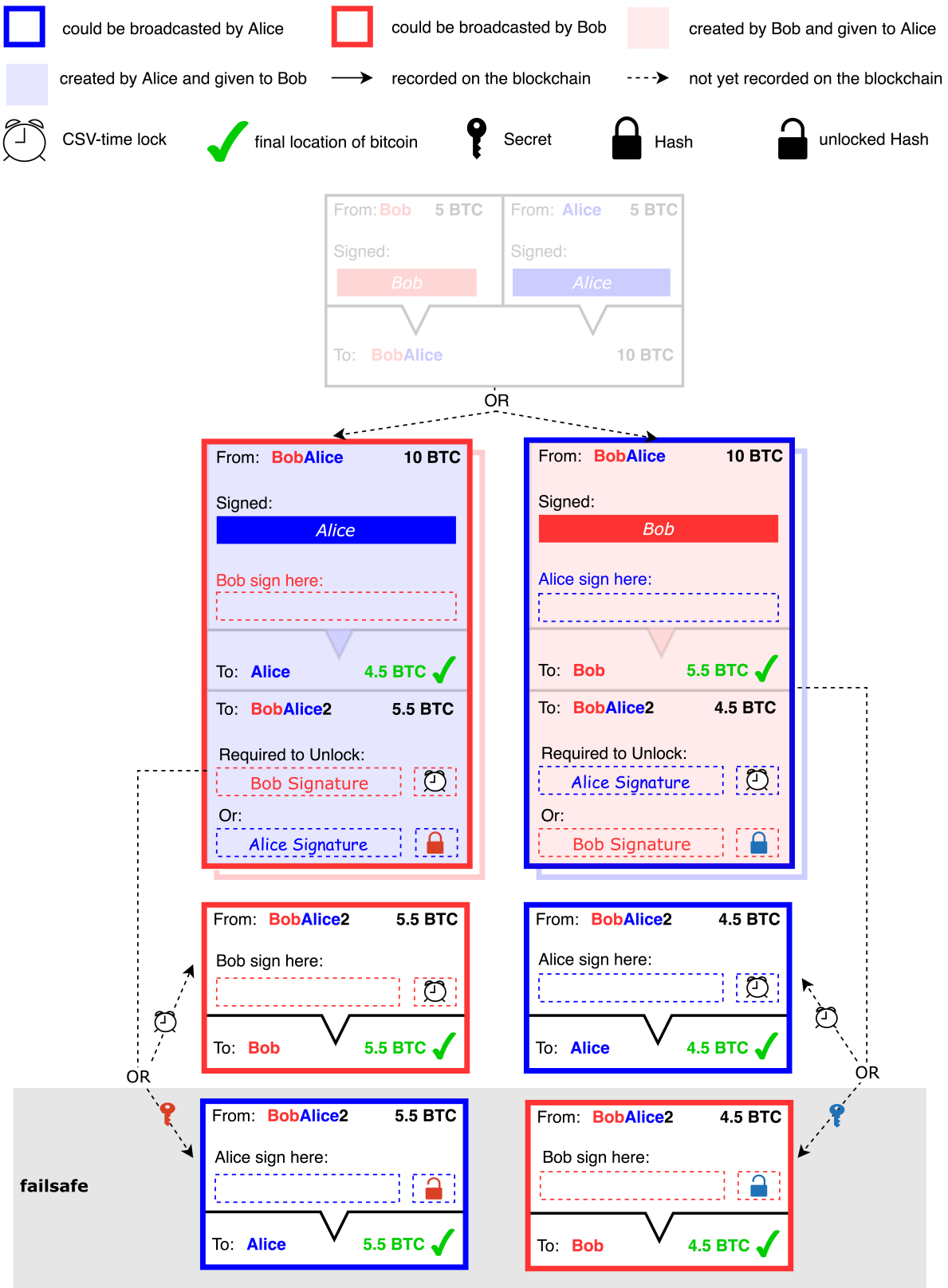


Figure 7.5: Update Transaction

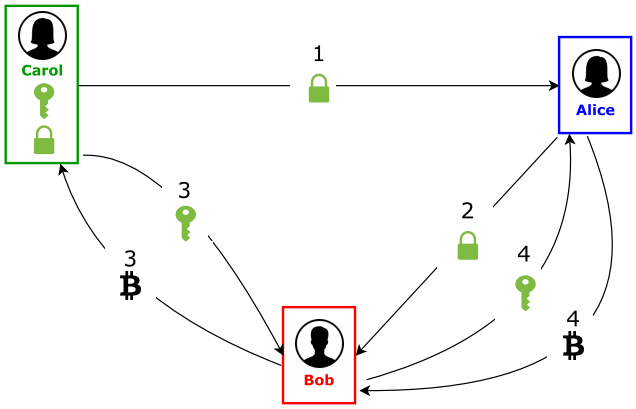


Figure 7.6: Multi-Party System

Figure 7.7: Bidirectional with HTLC

Bibliography

[1] Joseph Poon, Thaddeus Dryja: *The Bitcoin Lightning Network: Scalable Off-Chain Instant Payments*, January 2016. https://lightning.network/lightning-network-paper.pdf

[2] Grace Caffyn: *CoinDesk New Service Finds Optimum Bitcoin Transaction Fee*, CoinDesk, July 2015. http://www.coindesk.com/new-service-finds-optimum-bitcoin-transaction-fee/

[3] Scott Driscoll: *How Bitcoin Works Under the Hood*, July 2013. http://www.imponderablethings.com/2013/07/how-bitcoin-works-under-hood.html

[4] Aaron van Wirdum: *Understanding the Lightning Network, Part 1: Building a Bidirectional Bitcoin Payment Channel*, Bitcoin Magazine, May 2016. http://tinyurl.com/understanding-the-ln

[5] Aaron van Wirdum: *Understanding the Lightning Network, Part 2: Creating the Network*, Bitcoin Magazine, June 2016. https://bitcoinmagazine.com/articles/understanding-the-lightning-network-part-creating-the-network-1465326903

[6] Aaron van Wirdum: *Understanding the Lightning Network, Part 3: Completing the Puzzle and Closing the Channel*, Bitcoin Magazine, June 2016. http://tinyurl.com/understanding-the-ln-part3

[7] Robert McCone: *Ethereum Lightning Network and Beyond*, Arcturnus, October 2015. http://www.arcturnus.com/ethereum-lightning-network-and-beyond/

[8] Ethereum Foundation: *Ethereum Homestead Documentation*, Ethereum Foundation, 2016. http://ethdocs.org/en/latest/

[9] Paypal: *About Paypal*, Paypal, December 2016. https://www.paypal.com/ch/webapps/mpp/about

[10] Visa: *Run Your Business*, Visa, December 2016. https://usa.visa.com/run-your-business/small-business-tools/retail.html

[11] BitcoinWiki: *Transactioncost*, BitcoinWiki, December 2016. https://de.bitcoin.it/wiki/Transaktionsgebuehren

[12] Raiden Network 2016 : *Raiden.Network*, Raiden Network 2016, December 2016. http://raiden.network/

Chapter 8

Economical Impact of SDN and NFV for Telecom Operators

Lukas Braun, Jan Meier, Lu Da

This paper discusses about technical and economical benefits of Software-Defined Networking (SDN)and Network Functions Virtualization (NFV) in the environment of telecom operators that lead to a positive economical impact. Based on examples of both SDN and NFV we introducethe main concepts involved in these technologies, discussing their origin and their functioning. Both technologies have their advantages and disadvantages when contrasted to traditional no-software oriented networking. It is shown that the application of these principles reduces the development time, reduces maintenance effort and reduces the hardware overhead. That leads to a significant reduction of operational expenditure and capital expenditure. Therefore, in this paper we present examples on both SDN and NFV discussing their benefits and drawbacks related to the transition towards a software-defined and virtualized networking infrastructure. The transition from an old infrastructure to a new solution takes a lot of time and effort, as it is designed for a general purpose it is not very well suited for specialized, high demand, high performance applications (yet).

Contents

8.1 Introduction and problem statement

8.1.1 Traditional Networking

The Internet has become a vital part of our economy and our social lives. Its origins can go back to the ARPANET which was invented in the 60s and 70s of the 20th century [1]. The ARPNET first implemented the revolutionary concept of *packet switching*. Previous to this idea, the connection between two computers was thought as a constant connection, what is known as *circuit switching*. Packet switching introduced the idea to splits up the data into packets which are then sent individually. Along the path between two hosts the packets would pass intermediate stations which forwarded these packets based on a certain set of rules. These stations are called switches.
The ARPANET was designed as single big network with a uniform hardware landscape. This network soon grew and it became a network of networks. This was the birth of the internet as it is known today. The new topography no longer was controlled by a single organisation but by several, more or less independent ones. This required a new approach to define the standards in this network, away from a specific software implementation towards a more general communication protocol [1]. This lead today dominant Transmission Control Protocol/Internet Protocol (TCP/IP) which is still in use today.
At this time, introducing a central coordinator was avoided. This would have been a single point of failure and since the network infrastructure was not stable enough, too much of a risk. Therefore, the switches were designed to operate without any central control instance, to avoid a single point of failure [1]. Their routing decisions were made based on local knowledge of the network topology. The computational power back at this time was very limited and the switches were designed with special purpose hardware with special software. That resulted in devices which worked highly decentralized and with tightly coupled hardware and software.
As mentioned previously, the Internet's initial infrastructure consisted of end points and switches. Over the years more Network Functions (NF) were required and new device types were introduced. Devices implementing NF are called middle-boxes and were introduced for tasks such as protecting networks or deal with spikes in the network traffic. Today there is a huge amount of this specialized hardware produced by numerous vendors [2].

8.1.2 The flaws of the current system

This evolutionary process is not only evolving towards the creation of stable network infrastructure but also is leading to several shortcomings.
Changing the overall routing behaviour of a network is a difficult task in the current network infrastructure.
Optimizing the network infrastructure's over all routing behaviour to adapt to new requirements is a difficult task. As mentioned earlier the switching devices are designed to make their routing decision based on very limited knowledge of the over all network topology. This makes it hard change the over all behaviour of the routing in a network, because every device has to be configured individually. Today, the hardware landscape involves hardware from different vendors, which doesn't have a common configuration interfacer. This requires to tailor the solutions to the vendor specific products. Further, there is no central entity to control a network's devices. The devices must be configured via a remote connection of via physical access.
Another challenge of the current network infrastructure is to include new functionality such as the support for new protocols. This requires the collaboration of the hardware

vendor because the hardware and software of the networking devices is tightly coupled, what makes it hard for the customer to change it by them self. This also leads to longer development cycles because the new features have to be included in to the next generation of devices, what might take several years. This poses a major issue to companies which are acting in the fast evolving economy because they cannot react to market changes within reasonable time [3].
A similar issue is present for the middle-boxes, which are mostly purpose built hardware. They implement their functionality on closely integrated software, with stability and performance in mind. This makes it inherently difficult to change or extend their functionality. Again, dynamic changes are required by modern companies in order to react on short term market changes and reduce their development time of new products [2].
The middle-boxes are physical devices located in rack towers. They are integrated into the network by physical links. Changing their location in the network requires a physical displacement and a manual rewiring. These are both labour intensive and time consuming tasks especially for large scale data centers.
In this paper we are going to discuss how a combination of Software Defined Networks (SDN) and Network Function Virtualisation (NFV) can address these issues as well as what SDN and NFV can do individually. After presenting what these technologies are capable of we then show how these technologies can be combined and how they complement each other. After that, we will elaborate the economical benefits of a network infrastructure based on SDN and NFV for telecom operators and large scale server operators in general.

8.2 Background

8.2.1 Software Defined Networking (SDN)

8.2.1.1 Definition

The underlying idea of Software Defined Networking (in the following pages SDN) is to decouple the network control handling from network data packet handling. It was an attempt to move away from the simple approach of “Shortest Path Routing“ into a software managed environment. “In a software-defined network, a network administrator can shape traffic from a centralized control console without having to touch individual switches (...).“[4]

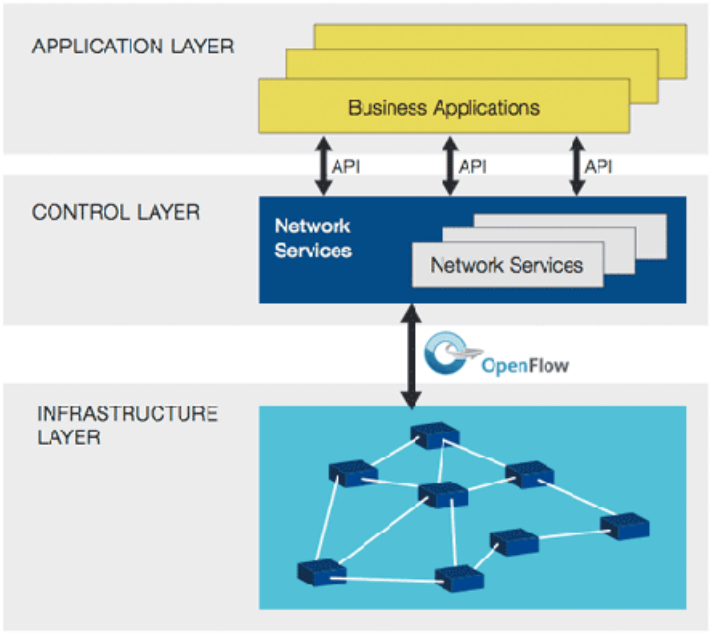


Figure 8.1: Initially from OpenFlow website [5]

As it is well shown in the graphic above SDN follows a three layer approach. The goal is to be able to program the network hardware with a central software that has the overview of the network. "SDN takes the control plane (how a network device will forward traffic) and separates it from the data or forwarding plane (a network device forwarding traffic based on the control-plane policy)" [6]. By doing so, a network engineer gains full control over any forwarding policies.

8.2.1.2 Benefits

Cost Reduction. SDN itself is just a piece software that is implemented on an existing network setup. Therefor, it is not a huge investment moneywise compared to a new hardware setup. There are even free solutions for SDN (Examples for SDN controllers: OpenDaylight, OpenContrail SDN controller, Floodlight open SDN controller, Ryu OpenFlow controller, FlowVisor OpenFlow controller [7]). By changing from a traditional network setup to a SDN it is possible to create network setups that can implement VLAN or layer 1 through layer 3 without expensive hardware.

Overhead Reduction. "In a physical environment, the isolation for the customer workloads requires configuring VLANs on separate networking devices, including routers, switches, etc. Since most of the networking is done at the SDN, it is easy for service providers to isolate the customer virtual machines from other customers by using various isolation methods available in the SDN.[8]"

Physical vs. Virtual Networking Management. Due to a central piece of software a need for a new networking device can be fullfilled very easily and fast. The deployment of new hardware, which is needed for new networking devices such as storage devices or servers, usually takes different stakeholders in a company to work together, which takes a lot of time." A virtual administrator can process the necessary changes without needing to collaborate with different teams[8]." Meaning that no real person is needed to be available

24/7 and there is no collaboration needed between different physical teams. As soon as a new networking device is needed the deployment itself is provided by a service.

Managing Virtual Packet Forwarding. Virtual packet forwarding compared to normal packet forwarding gets easier as an administrator does not need to access a specific machine on a specific hardware. One might not be able to provide *e.g.*,internet access in a specific bandwidth to a virtual machine and be able to change that setup within seconds when needed.

Reduced Downtime. By viratualizing an entire network setup there are no critical nodes any more. *E.g.* there is no critical network switch that leads to a complete system breakdown when failing. By doing so upgrades can be done without risking a complete system failure. The central controller of SDN also supports the creation of snapshots of the current setup which adds an additional layer of failur protection. This is not possible in a classic setup, as the preferences of each device needed to be saved and reapplied afterwards.

Extensibility. "Since SDN is software-based, it is easy to use SDN API references for vendors to extend the capabilities of an SDN solution by developing applications to control the behavior of networking traffic.[8]" The usage of Software Defined Networking in a network setup is therefore not limited by what it is capable of out of the box. It is possible to extend its possibilities throughout time and really have a tailor made solution for a specific setup.

Central Networking Management Tool. Different hardware mostly has different software management tools. Reconfiguring or managing physical networking hardware can therefore be a very time consuming task. By centralizing the management of hardware it simplifies this task by a lot.
Having the possibility to monitor the network, storage and computing needs of a network setup at any given time, a comprehensive Information Technology (IT) strategy can be worked out much more precisely. Resellers gain the ability to plan possible IT strategies better than before and the upcoming costs of a change in a setup can be planned much better.

8.2.1.3 Challenges

Security. One of the main issues associated with SDN in the last couple of years clearly is security. As it is mentioned in "Are we ready for SDN? Implementation challenges for software-defined networks" from 2013; "Potential security vulnerabilities exist across the SDN platform. At the controller-application level, questions have been raised around authentication and authorization mechanisms to enable multiple organizations to access network resources while providing the appropriate protection of these resources" [9]. One part of a potential solution to the problem is the implementation of rolebased authorization as the actual requirement of SDN should be isolation of applications and resources.

Performance. "General-purpose processors (CPUs/GPPs) provide the highest flexibility. High-level programming languages and design tools enable the highest design abstraction and the rapid development of complex packet processing functions.[...] Network flow Processors (NPUs/NFPs) are optimized processor architectures for network processing. Instructions and interconnects are tailored for processing packetized data.[...] Application-Specific Standard Products (ASSPs) are the cornerstone of high-performance networks.

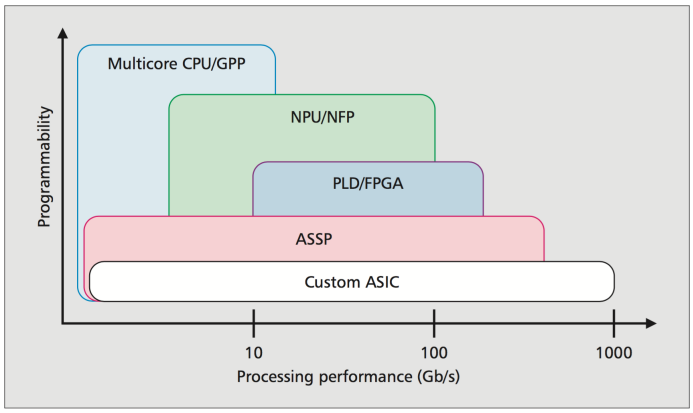


Figure 8.2: Networking processing: performance vs. programmability [9]

They are designed and optimized for widely used functions or products aiming for high volume. The drawback of ASSPs is their limited flexibility. (Custom) Application-specific integrated circuits (ASICs) are proprietary devices custom-built by system vendors (*e.g.*, Cisco, Huawei, Juniper) when standard products are unavailable and programmable solutions are unable to meet performance constraints.“ [9] That really shows the potential application for SDN. SDN products are seen as the compromise of processing performance and programmablity with a strong tendency towards programmablity. It can be seen where the limitations to a SDN solutions are. As soon as great performance is needed other hardware based solutions are needed.

8.2.1.4 Examples

As it is shown in the previous paragraphs about SDN, it seems to be a neat solution for an existing problem with near endless applications. But whithout an example there is no chance the reader will understand what the benefits really are.
In January 2014, COHO DATA magazine published an ariticle about “Data, Storage, and SDN: An Application Example“ [10], which shows a real and easy to understand problem and solution of SDN. The problem given was about the random I/O of hard disks. The performance of modern hard disks (or harddisks in general) drops to significantly as soon as search is required on the disk (“to 2 MB/s or less “) [10]. The trend towards virtualization lead to more and more virtual machines running on servers “(virtual harddisks, instead of individual documents), broadly known as “the I/O Blender“ effect“[10]. The bottleneck to writing speeds in networks has always been the hard drives not the connection itself.
Early Solid State Discs (SSD) gave a significant increase in random access to the storage servers but behaved very similar to HD‘s under constant read write tasks, leading to the usage of SSD‘s as a cache. In 2010 PCIe-based flash devices arrived and changed the way storage was planned and executed.
“This is one of the predicating observations that we made a few years ago in starting our company: That storage was about to change fundamentally from a problem of aggregating low performance disks in a single box into a challenge of exposing the performance capabilities of emerging solid state memories as a naturally distributed system within enterprise networks. By placing individual PCIe flash devices as addressable entities directly connected to an SDN switch [...][10].“

The Single IP Endpoint. NFS, one the most common storage protocols, assumes that the server has a single IP address. That is not the case in modern network setups. The integration of SDN in the form of *e.g.* the OpenFlow software allows the network to assign

the that very connection to a lightly loaded node in the network. The software controlling that connection is able to change the path at any given time and modify between storage clients and storage resources.
“This decoupling of client connections from a specific storage controller at the end of the wire solves an immediate scalability problem that until now has needed either interface changes on the client [...] or complex administration [...]. “ [10] Data is therefore treated as a completely fluid resource.

8.2.1.5 Technical Approach

To show what the technical change of SDN in comparison to conventional networking is, one must take a look at the different planes involved in a telecommunication architecture. ”The control plane, the data plane and the management plane are the three basic components” [11], that are implemented in routers and switches. ”In conventional networking, all three planes are implemented in the firmware” [11] of the given hardware.
”The data plane (sometimes known as the user plane, forwarding plane, carrier plane or bearer plane) is the part of a network that carries user traffic.(...)Data plane traffic travels through routers, rather than to or from them.” [11].
”The control plane is the part of a network that carries signaling traffic and is responsible for routing. Control packets originate from or are destined for a router. Functions of the control plane include system configuration and management” [12]
The management plane is a part (sometimes considered) of the control plane and transports administrative traffic.
What a SDN approach means is to seperate the control plane from the data plane. A centralized instance has an overview over the network and has the ability to modify routing behaviour depending on the current network demand.

Northbound ”In a software-defined network (SDN) architecture, the northbound application program interfaces (APIs) are used to communicate between the SDN Controller and the services and applications running over the network. The northbound APIs can be used to facilitate innovation and enable efficient orchestration and automation of the network to align with the needs of different applications via SDN network programmability.[13]”

Southbound The Southbound or southbound API (application program interface) are used to handle the communication between the hardware (switches and routers of the network) and the SDN Controller. Maybe the best known southbound API is OpenFlow, that managed to establish an industry standard. The advantage given by such an application program interface is the ability to change the routing behavior of switches and routers in real time according to the current demand of the network.

8.2.2 Network Functions Virtualization (NFV)

8.2.2.1 Definition

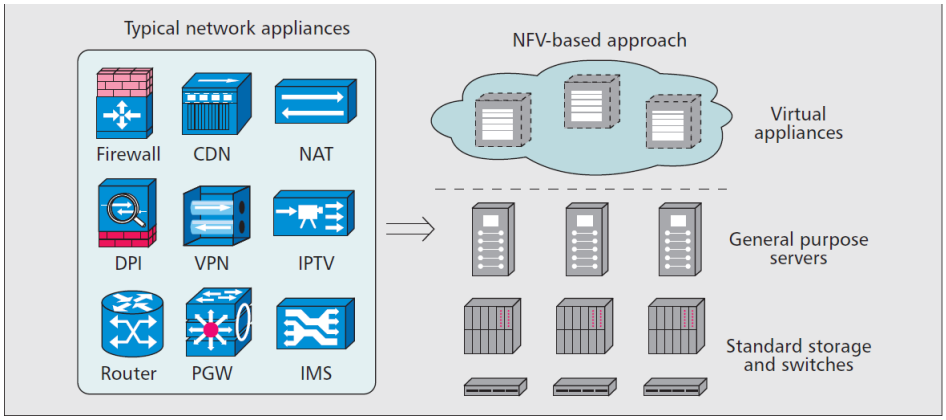


Figure 8.3: "From dedicated hardware-based appliances for network services, such as firewalls, Content Delivery Networks (CDNs), Network Address Translation (NAT), Deep Packet Inspection (DPI), Virtual Private Networks (VPNs), IPTV, routers, packet data network gateways (PDN-GWs or PGWs)"
[14]

"Network Functions Virtualization (NFV) is an initiative to virtualize the network services that are now being carried out by proprietary, dedicated hardware"[15]. Therefor as it is shown in the graphic above services that currently need dedicated hardware run on general purpose server and standard hardware. The main goal of NFV is to decouple network functions and services from its hardware. It is a step towards a virtualization of these services on VM (Virtual Machines). From an architectural perspective, NFV are pieces of software with networking capabilities deployed on a general-purpose hardware. Instead of setting up a new router or a new switch, a new virtual machine is deployed that runs on general purpose hardware and virtualizes the service required within second, if not less. Scalability is endless and a change in demand for a specific service is no longer and issue as well as the possibility to handle overhead as well as possible.

8.2.2.2 Benefits

Scalability. Ever changing user needs created a need for a responsive and quick network setup. The possibility to monitor a bottleneck and quickly change the resources for a service is key for the services users need today.

Cost. Standard server hardware is much cheaper than specialized hardware which is needed to deploy network services on non NFV-based networks.

Flexibility. A new service can be tested and deployed much faster than on standard proprietary hardware. The demand of a quick and fast service deployment has increased by a tremendous amount in the past couple of years.

Security. " In a cloud environment, multi-tenancy requires virtual resources to be logically separated among tenants. Using orchestration, certain VNFs can be deployed on separate compute nodes, and they can be further segregated by using separate networks. In addition, using security zones allows VNFs to be deployed on hosts that satisfy security-pertinent criteria, such as location and level of hardening (*e.g.*, some hosts may employ the trusted computing technology). [...]NFV can reduce the operational impact of deploying security updates. An upgraded instance of the VNF can be launched and tested while

the previous instance remains active. Services and customers can then be migrated to the upgraded instance over a period of time (the length dictated by operational needs). The older instance with the un-patched security flaw can be retired once this is complete“ [16].

Improved innovation cycle The way an improvement or change is done by today's standards still relies on proprietary hardware. The new service then needs to be implemented in some kind of software on this very hardware. This process is very time consuming, which has a lot of interfaces and very costly. Instead of running it on a proprietary hardware you can go ahead and deploy it on generic server style hardware and virtualize all the functions needed. It is also believed, that all the generic hardware needed is available today and doesn't need to be developed first.

8.2.2.3 Challenges

Lack of Control. One of the main challenges for todays NFV is the lack of control. Taking the example of switches [17], there is a significant loss in performance (compare next paragraph) in direct comparison to running a switch on bare metal. This is partly due to the higher performance achieved by specific hardware for the purpose and partly due to the lack of control. “But NFV poses a problem that might be beyond the OVS team‘s scope. Namely: to get the most out of a virtual switch, you need a lot of low-level configuration - and that stuff gets abstracted away by OpenStack, Taylor says. The lack of fine-tuning limits the amount of performance you can wring from the switch.“[17]

Performance. As a result of the lack of control and other factors, performance is still one of the key issues that you may face when using NFV.

8.2.2.4 Examples

Described in the sections above, NFV is a key method for resource optimization in small and medium sized networks without the need of hyper-performance. But simply stating the virtualization part of NFV do not illustrate what is it- an example is needed.
One of the best illustrating examples is a firewall. “A firewall is a network security system, either hardware or software-based, that controls incoming and outgoing network traffic based on a set of rules“[15]. In theory, a firewall will decide weather in incoming traffic will be forwarded to the requested Port and IP address based on the rules that it was given. It does not matter if that task is completed by software or hardware unless great performance is needed (in that case hardware is far superior). The implementation of a firewall as a virtual network function delivers the same functionality but it is running on a virtual instance. The usage of NFV in that case leads to a more efficient usage of the hardware involved. A minimized usage of middle-boxes, which are expensive and require maintenance leads to a significant cost reduction.

8.3 SDN and NFV

8.3.1 Virtual Network Infrastructure

Network services are services which are accessible through the internet. They range from services offered to other applications, such as database system or registry systems, to services designed for human beings such as the World Wide Web or movie streaming. These services require a infrastructure to provide a stable service, which is called network infrastructure. As mentioned above, this infrastructure is traditionally built from physical

devices linked by physical links. However, SDN and NFV allows to create Virtualised Network Function Infrastructure (VNFI).

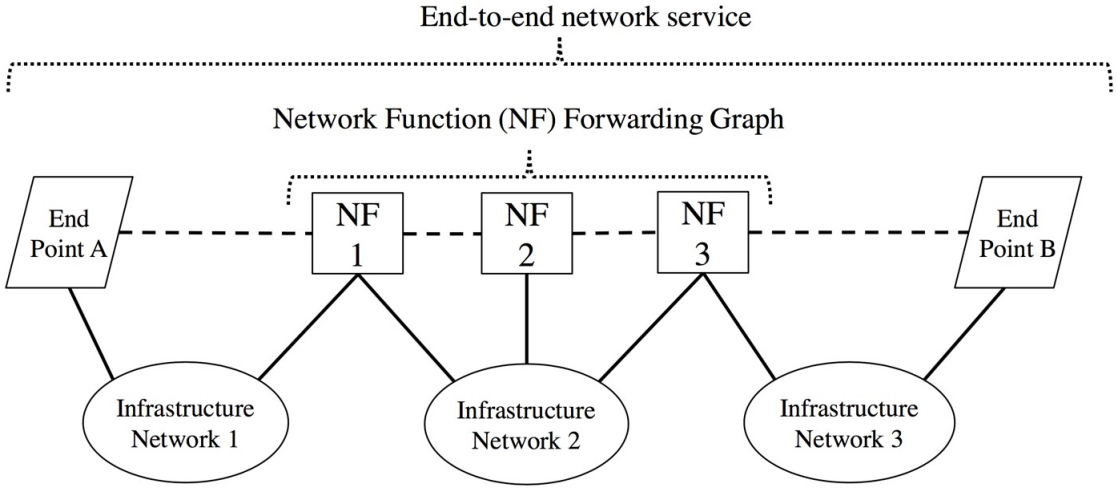


Figure 8.4: Traditional network service [18]

Figure 8.4 shows a schematics of a network function. For example, this could be a online movie streaming service. On the left hand side, there is a client which could be a Smartphone, Laptop or SmartTV. In the middle, there is a collection of three Network Functions that are connected by links. The functions and links together form the Network Function Forwarding Graph which defines the logical structure of the network service's single components. For the streaming example we assume that NF 1 and 3 are Load Balancers and NF 2 is a firewall. On the right hand side there is the endpoint, which provides the actual service. In our example this would be the movie content provider.
A realization of this service using VNFI is found on figure 8.7. It shows how the middle part, consisting of the various Network Functions, is realized using virtualized hardware resources. The Forwarding Graph broadly remains unchanged. There are still three Network Functions which are linked sequentially. However, the virtualized version of NF 2 consists of several sub VNF instead of a single monolithic one. The links between the single VNF are logical links and might not correspond to physical ones and the VNF might be distributed among several physical devices. This is depicted by the dashed arrow from the lover half of the picture to the upper half. Between the physical and logical representation of the Network Service a virtualization layer is found. This virtualized layer abstracts the physical resources away and makes them transparent to the VNF.

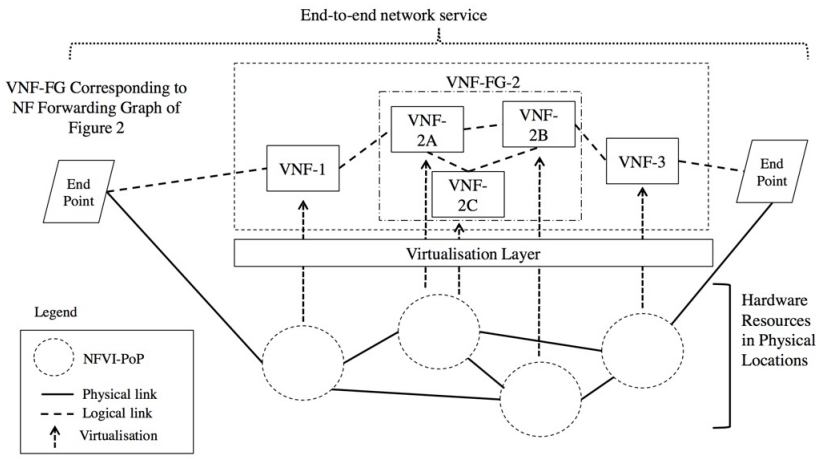


Figure 8.5: SDN and VNF enabled network service [18]

Figure 8.6 shows the VNF reference architecture as proposed by the European Telecommunication Standards Institute (ETSI). It consists the three components: NFV Management and Orchestration, Virtual Network Infrastructure and Virtual Network Functions.
The sub components of the *NFV Management and Orchestration* part are responsible for the management of the various resources. The *VNF Manager* is responsible for the life cycle management of the NFV instances. For example, the VNF Manager mounts more instances of a certain VNF in a high deamnd situation and will shut down some if the demand decreases. The *Virtualized Infrastructure Manager* manages the hardware components and integrates them to a transparent virtualised platform. For instance, if a new storage storage capacity or more computational power is added to they are connected to the existing resources and newly added network links are added to the physical network infrastructure. The *Orchestrator* delegates tasks to the previous two components in order to create a certain Network Service. To to do so, it uses the functionality of the VNF Manager and Virtualized Infrastructure Manager. What the Orchestrator exactly sets up is defined in the *Service, VNF and Infrastructure Description*
The *NFVI* component deals integrates all the hardware and software which is required to deploy the VNF. The components are integrated into virtual instances such that the VNF can run without any awareness of the physical devices. The hardware may or may not be distributed among several locations. In either case the NFVI component deals with the tasks, such as integrating the hardware and setting up the physical networking infrastructure to, to create a consistent platform.

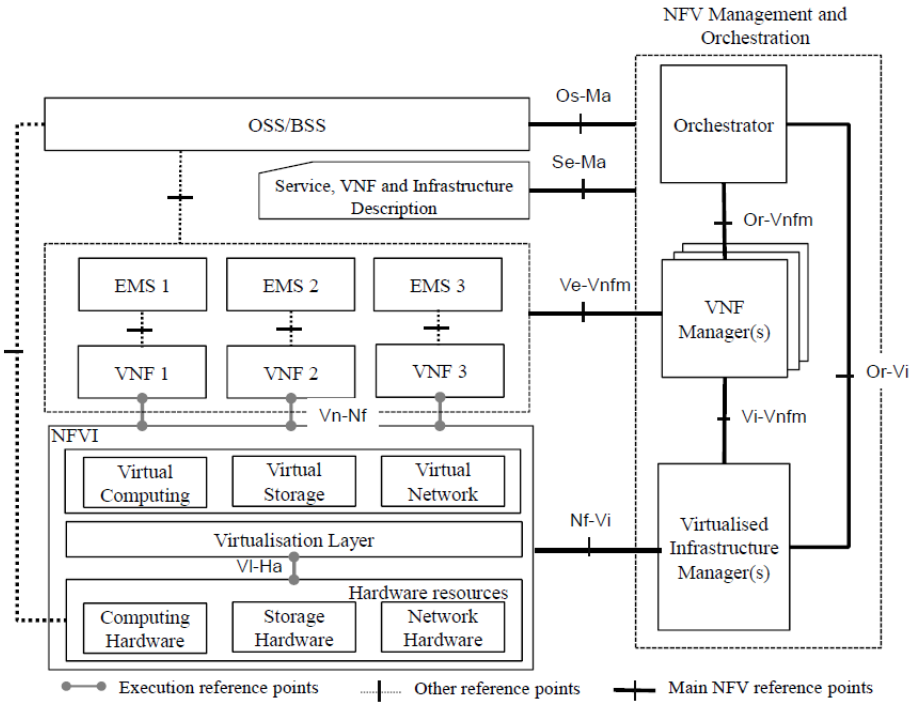


Figure 8.6: VNF reference architecture proposed by the ETSI.[18]

The *VNF* Component is the location where the actual VNF are running. A single VNF might be composed from several sub VNF.
So far, software defined networks has not been mentioned specifically. In a early version, the NFV reference architecture has been design without specifically referring to SDN. In fact, the whole framework is basically feasible without. But it is apparent, that SDN heavily increases the dynamic capabilities of the NVFI, especially when the physical resources are geographically separated.

8.3.2 Advantages of SDN and NFV enabled network Infrastructure

The SDN and NFV enabled infrastructure reduces the development time of new products [2]. On the one hand implementing new services becomes less dependent on the hardware vendors. The required changes to the hardware can be implemented by the developer. These changes no longer depends on changes to the vendor specific hardware which often have development cycles up to several years. Furthermore, the deployment time for new services can be reduced. The roll-out of the new functionality no longer requires new hardware but merely mounting new virtual functions, which provide the new service. Also, the integration of the new functionality no longer require time consuming manual changes to the network topography. It can be done on a central entity using the functionality of SDN.
A second benefit are the reduced maintenance costs. Network maintenance in a large scale data center can easily occupy several network administrators [19]. With SDN and NFV many of these tasks can be automated or done in a more efficient manner requiring less workforce.
Further, SDN and NFV increase the scalability of networks and services [2]. During high demand periods more instances of a certain NF can be mounted and integrated in the existing network. During low demand these instances can be shut down. The integration of these instances is done in a completely autonomous manner by using SDN. This way the resources can be tailored to the demand on a very fine granular scale, minimizing the power consumption of the network infrastructure.

8.3.3 Network Virtualisation and IoT

A use case where SDN and NFV will be useful is the network infrastructure for the numerous Internet of Things (IoT) products which will appear on the market in the coming years. The use case we present here is based on a paper by Omnes et all [20].
The concept of IoT relates to every day objects, such as cars, home electronics and industry Iachines, which will be connected to the internet. What in 2014 already included about 1.5 billion objects is expected to increase to 70 billion by 2020 [20]. These IoT devices will belong to a wide variety of different domains including personal gadgets, medical equipment and vehicles.
This new type of devices will fundamentally change the characteristics of the internet traffic. The rapidly increasing number of devices will massively increase the amount of end points in the internet. Also, the mobility of these endpoints will increase. This boosts the dynamics in the mobile the traffic patterns. Also, characteristics of the endpoint's hardware will change and the over all diversity will increase. Traditionally most endpoints were fully powered computers with sufficient computational power for all tasks networking requires. Many of the new IoT devices will be small, battery powered and energy efficient devices. These characteristics will require energy efficient and computational inexpensive protocols for the network communication.
Over all we can say that the internet traffic is going to be less predictable. Therefore, the internet infrastructure has to become more flexible, agile and scaleable in order to cover all eventualities. Today, mostly dedicated hardware is used in mobile communication infrastructure. Under the aspect of an increasing demand for flexibility that would be a hazardous choice for future infrastructure choice. Promoting a system which founds on SDN and NFV is a more reasonable choice, since it will be flexible and dynamic and can deal with changing requirements.
As the IoT trend will continue many domains of our economy, which formerly didn't rely on internet communication, will start to do so. This will trigger a vast demand for new

products and yet unknown services. As stated earlier, the development of new services is difficult when it relies on traditional hardware which is hard to extend and configure. A network based on virtualisation techniques on the other hand will enable a broad variety of new products for this new domains. Also, the development time will be reduced.
As an illustrating example, we will take a closer look at the IoT gateways in a domestic setting. Every IoT device will require some sort of gateway to unfold their full potential. These gateways have to cover different communication such as Bluetooth, WiFi or Ethernet in a transparent manner for allowing the real IoT experience. They also will allow the devices to trigger certain actions or connect to the internet. Figure 8.7 shows how such an infrastructure might look like.

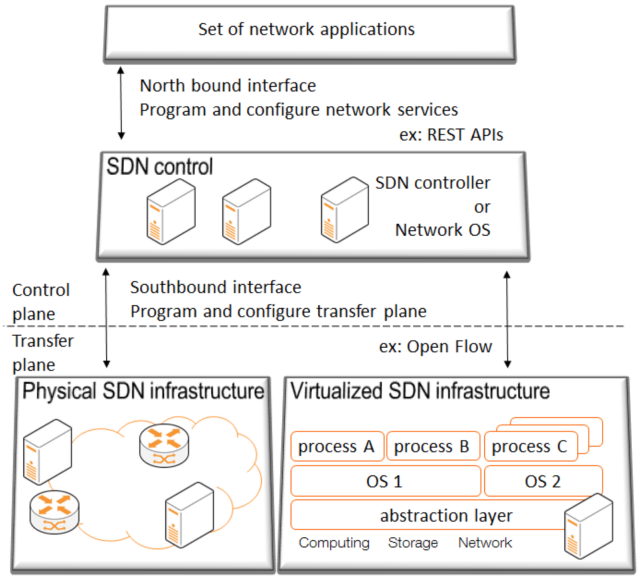


Figure 8.7: SDN and NFV enabled network infrastructure for IoT [20]

These gateways could furthermore implement additional local storages, security features or run applications from other domains. Now the question arises where these services should be located. They could be located in the gateway it self or cloud of a trusted provider. Both options have their advantages and drawbacks. A implementation in the gateway will have a high availability and short response time, because it is located close to the devices. On the other hand, the gateway will have limited computational power. High performance components are expensive, and installing them in every gateway, to cover the rare events where high computational power is required, is economically infeasible. Putting the services into the cloud allows the shared usage of high performance hardware. But placing the services further away from the IoT devices will increase the response time. Therefore a hybrid approach is most favorable since it combines the advantages of both solutions.
The drawback of a hybrid approach is that the implementation will require a more sophisticated control logic. This logic would be very difficult to implement with traditional networking infrastructure. But with SDN and NFV this implementation would be feasible.

8.4 Potential Economical Improvements

With the fast development in recent years, SDN and NFV have shown great potential in cost reduction, efficiency improvement and flexibility enhancement, compared to traditional networking system. According to IT industry body NASSCOM, the global market of " Internet of Things" (IoT) will reach 300 billion US dollar in 2020, and SDN as well as NFV will play critical role in the expansion of the IoT market [21]. The Information

Handling Services (IHS)s report shows that with a 86 annual growth rate from 2015 to 2020, the combined market of SDN and NFV is expected to touch 45 billion US dollar in 5 years [22]. Telecom operators are shifting their focus from traditional networking system to SDN and NFV. What are the reasons behind this? Some economical improvements play vital roles in the transformation.

8.4.1 CAPital EXpenditure (CAPEX)

Nowadays, the CAPital EXpense (CAPEX) accounts for a major part of the costs that telecommunication operators face today. All the towers, 4G networks, radios and trucks are included in this area. To deal with the large volume of traffic, telecommunication providers build large data centres, which occupy large amount of space. The servers consume a huge amount of electricity every day and cooling also need to be provided to maintain function of the whole physical system. All of this requires tremendous amount of energy and is considerable amount amount of capital and operational cost of these companies. By deploying SDN and NFV technology, telecommunication operators may be able to save a large percentage of expenditure. For example, "Google were one of the first companies to announce it had moved its data centres and Wide Area Networks (WANs) to SDN" [23]. The computing performance of their data centres is claimed to improve by 60% to 70% in WAN utilization and therefore, the costs are reduced sharply at the same time.
Verizon is a major American telco provider. By July 2016, they had 142.5 million subscribers and surpasses any other U.S telecommunication providers [24]. According to Khan, Farhan Ahmad's paper Virtualized EPC: Unleashing the potential of NFV and SDN [23], in 2012, Verizon's capital expenditure was $ 9 billion. To establish it's 4G LTE network, Verizon expects to spend $4 billion a year in the following couple years. The core network equipment is estimated to be $ 500 million. By deploying virtualised network network infrastructure with commercial IT equipment, the company expects to save 20% of their expenses, which is approximately $100 million.
Not only large enterprises will reduce lots of costs, new started-up will also benefit from SDN and NFV. Virtualized network technology helps them save money in capital investment and allow them to use to invest money in other critical fields. It "has eliminated the capital barriers to computing resources, or at least it has drastically decreased the threshold to obtain the benefits of having them" [25]. In the book " SDN and NFV Simplified: A Visual Guide to Understanding Software Defined Network and Network Function Virtualization" [25], Jim Doherty illustrate this many vivid examples. Two college students Jeff and Carol want to design an application about helping people settle arguments. However, they can not afford the cost of purchasing server and operating system, which will cost them $5000 and the monthly hundreds of dollars Internet contract fee with an Internet service provider (ISP). With SDN and NFV, however, what they need to do is renting computing resources with only $30 per month. They also avoid expense Internet connection fee because it is not fixed amount but based on traffic volume they currently have. If they need to build additional bandwidth, they can also do so in anytime.

8.4.2 OPErational EXpenditure (OPEX)

Apart from reducing capital expenditure, SDN and NFV also fundamentally change traditional carrier operating model. With centralized and virtualized networks, operators do not have to invest much capital on equipment support and maintenance. The management of the whole networking infrastructure will also be simplified and automated. Therefore, the overall maintenance costs for the operational system will be considerably diminished.

In 2015, Arthur D Little, an international management consulting company from America, and famous Nokia Bell Labs conducted a study that examined the economical impact of network virtualization on telco operators in 35 European countries' telecommunication [26](Arthur D little & Bell Labs, 2015). According to the report, in 2013 those operators had 250 billion euros adjusted revenue in total and 150 billion euro annual operational expenditure. The researchers calculated that the impact of onboarding SDN and NFV would be worth 14 billion euro every year. Which would be equal to 10% of annual operational expenditure. The picture below showed that the operational cost could be reduced to 111 billion annually, which was equal to 16% of total cost, by further reducing cost of different functions in the operational model.

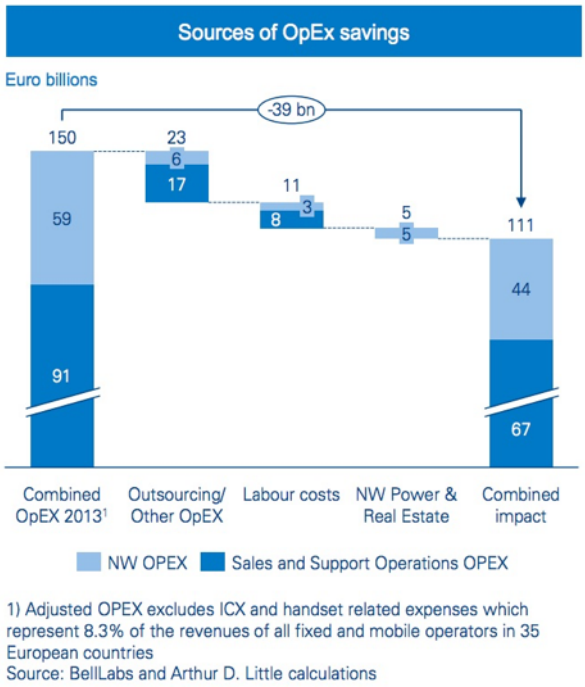


Figure 8.8: Sources of Operational Savings [26]

These two companies conducted the study in two different aspects: technology onboarding and revised operating model. On one hand, technology onboarding examined the impact of transforming from traditional networking system to SDN and NFV. On the other hand, revised operating model referred to "taking advantage of flexible new technologies to streamline operations in the business layer from marketing, sales, back-office and associated IT" [26]. Traditional networking systems have developed a complicated vertical pattern, which "lacks of unifying services and resource abstraction, and a bloated back-office support environment that includes layers of legacy systems, applications and expensive IT workarounds" [26]. The flexibility of SDN and NFV allows operators to includes new capacities into existing systems without adding additional physical devices.

8.4.3 Efficiency and Utilization

In addition to CAPEX and OPEX, efficiency and utilization improvement is another great advantage of SDN and NFV. In static server models, companies usually need to set up several servers to run different applications, which is called sever proliferation [25]. On average, only 10% to 15% of the servers capacity will be utilized. If a company decides to run a new application, it needs to purchase another sever, even though other severs have free capacity. Network virtualization can solve this problem by using centralized servers

in the cloud. This will largely increase the utilization of servers and thus lead to a positive economic impact.

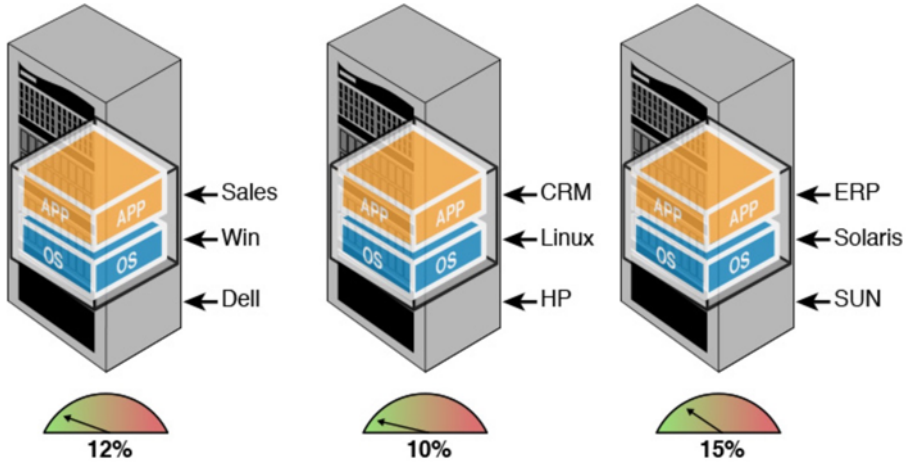


Figure 8.9: Static Server Model
[25]

Our report mentioned above that on average the utilization rate of non-virtualized serves is only 10% to 15 %. Jim Doherty explains the reason in his book [25]. Many companies' data centres have a usage pattern like shown in figure 8.10. That is because the traffic volume is circulated and increased time by time and companies have to set up the minimum server capacity to meet the expected maximum traffic volume. That leads to the terrible inefficient usage of servers. The next diagram shows the overhead cost after adding other capital expenditure.

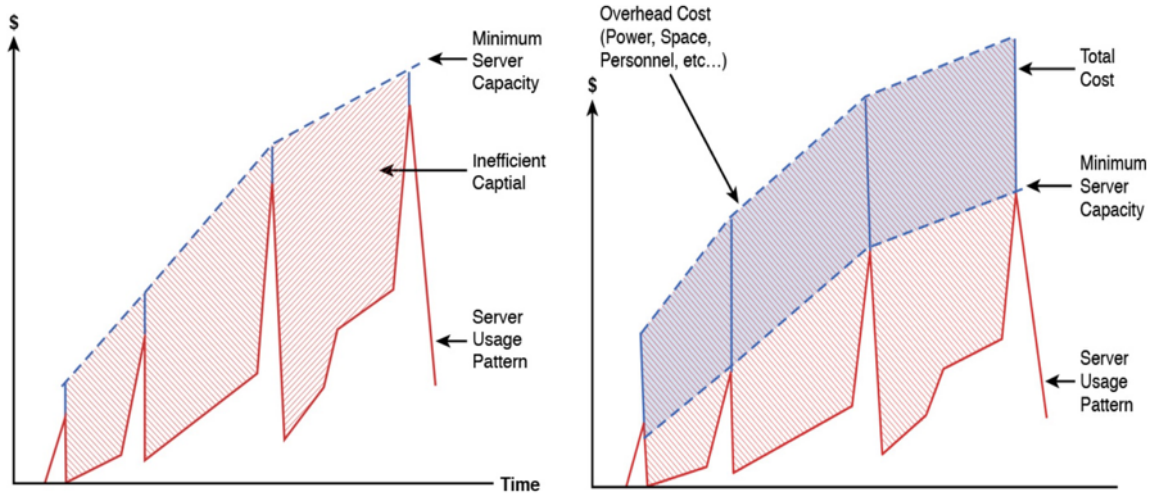


Figure 8.10: Usage pattern of Static Server Model
[25]

However, companies still earn profit with such a model. Although the cost goes up all the time, companies also make money proportionally at the same time. Therefore, they can set up more data centres and attract more traffic volume. But the cost will surpass the benefit as soon as the technology reaches the point of diminishing return. Fortunately, SDN and NFV will solve this problem. Companies can increase servers' capacity according to their need. Now, it becomes instantaneous serve capacity instead of minimum serve capacity. Companies can also get rid of those overhead costs at the same time. Therefore, SDN and NFV help operators save a lot of money and eliminate inefficiency at the same time.

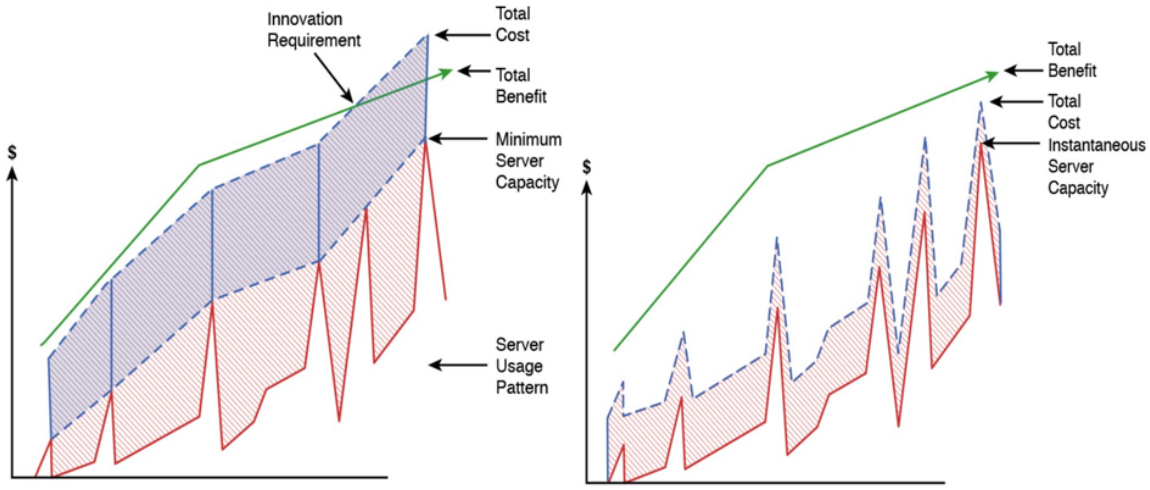


Figure 8.11: Comparation of usage pattern of Static modle and Virtualization model [25]

8.5 Conclusion/Going forward

The domain of network infrastructure is currently undergoing a fundamental change. The new technologies SDN and NFV will finally be used by all major network providers. The transition from traditional network infrastructure to the new virtualised one will still take some time. Replacing the existing hardware is a difficult task and will cost a considerable amount of money. But the advantages, economical and operational, of the new technologies are too significant to not use them.

We think that SDN and NFV will have the biggest economical impact on medium and large companies. Major telecomm players would implement that on their small/medium data centers, such as the Central Offices (CO). Further, specialized hardware, which will still be required by large specialized companies, still can be included in the virtualised hardware landscape. Medium companies benefit from SDN and NFV from reduced over all hardware costs and better utilization. Before the introduction of NFV, the network infrastructure required many different devices. This was a costly investment and required also a considerable amount of space. Also, the infrastructure was often not very efficiently used, because the traffic did not exhaust the devices capacities but it still needed to be dimensioned for the maximum traffic expected. Virtualised infrastructure has the advantage that the functionalities run on general purpose hardware and the resources can be assigned to the tasks on demand. Therefore, the same hardware will be used to tackle spikes in demand of the different functionalities. Thus, the overhead of devices which are not used very heavily will be reduced. Also in case of insufficient resources, adding a single general purpose machine will increase the performance of all network functionalities.

We do not think that small companies will directly benefit a lot from this new technology. Small companies often keep going with cheap off-the-shelf products, sufficient for their straight forward usecases. On the other hand the tendency to outsource the infrastructure to the cloud, will lead to more companies with no operational network infrastructure at all. But since the cloud providers use SDN and NFV they will benefit indirectly with lower costs and higher performance.

The private customer will benefit from the new technology as well. We are experiencing a transformation that more and more aspects of our every day life is entangled with the Internet. At work, we depend on communication applications or online collaboration platforms, which are infeasible without the Internet. In our private time we spend a considerable time communicating with each other or consuming services, such as YouTube or Facebook, which are driven by the Internet. Therefore, we are interested in infrastructure

which runs smoothly and can adapt to an increasing demand. SDN and NFV will help the companies and telcom providers to guarantee such a service in a reliable and safe fashion. The private customer, but also the business, will benefit from faster product release cycles. SDN and NFV simplify and accelerate the development of new software. Therefore, solutions to new problems will be available quicker. This especially will be interesting because of the trend towards IoT. These Internet enable object will drastically increase the internet capacity demand and introduce new quality features. For example, a IoT enabled production line of a car manufacture strongly depends on a reliable network with low latency. Otherwise the production will stop and cause a huge monetary loss. SDN and NFV are well suited to address these issues and create an infrastructure that meets the specific demands of specific situations for a reasonable amount of money.

Bibliography

[1] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1617–1634, 2014.

[2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, 2015.

[3] O. N. Fundation, "Software-defined networking: The new norm for networks," *ONF White Paper*, 2012.

[4] J. Burke, "software-defined networking (sdn)," 2013. [Online]. Available: http://searchsdn.techtarget.com/definition/software-defined-networking-SDN

[5] ——, "What is SDN? The answer now includes automation and virtualization," 2015. [Online]. Available: http://searchsdn.techtarget.com/tip/What-is-SDN-The-answer-now-includes-automation-and-virtualization

[6] E. Banks, "Sdn basics: Understanding centralized control and programmability," 2014. [Online]. Available: http://searchsdn.techtarget.com/tip/SDN-basics-Understanding-centralized-control-and-programmability

[7] M. McNickle, "Five must-know open source sdn controllers," 2014. [Online]. Available: http://searchsdn.techtarget.com/news/2240225732/Five-must-know-open-source-SDN-controllers

[8] N. Sharma, "Eight big benefits of software-defined networking," 2015. [Online]. Available: http://www.serverwatch.com/server-tutorials/eight-big-benefits-of-software-defined-networking.html

[9] B. Fraser, D. Lake, C. Systems, J. Finnegan, N. Viljoen, and S. O. E. N. Etworking, "Are we ready for sdn ? implementation challenges for software-defined networks," no. July, pp. 36–43, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6553676/citations

[10] A. Warfield, "Data, storage, and sdn: An application example." [Online]. Available: http://www.cohodata.com/blog/2014/01/22/data-storage-and-sdn-an-application-example/

[11] M. Rouse, "data plane (dp)," -. [Online]. Available: http://searchsdn.techtarget.com/definition/data-plane-DP

[12] ——, "control plane (CP)," 2016. [Online]. Available: http://searchsdn.techtarget.com/definition/control-plane-CP

[13] SDXcentral, "What are SDN Northbound APIs?" 2016. [Online]. Available: https://www.sdxcentral.com/sdn/definitions/north-bound-interfaces-api/

[14] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, 2015.

[15] M. Rouse, "Network Functions Virtualization (NFV)," 2016. [Online]. Available: http://searchsdn.techtarget.com/definition/network-functions-virtualization-NFV

[16] A. Lemke, "How to manage security in nfv environments," online, 2014. [Online]. Available: https://insight.nokia.com/how-manage-security-nfv-environments

[17] C. Matsumoto, "Nfv performance should be a bigger issue." [Online]. Available: https://www.sdxcentral.com/articles/news/nfv-performance-bigger-issue/2015/01/

[18] G. ETSI, "Network functions virtualisation (nfv): Architectural framework," *ETSI GS NFV*, vol. 2, no. 2, p. V1, 2013.

[19] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[20] N. Omnes, M. Bouillon, G. Fromentoux, and O. Le Grand, "A programmable and virtualized network & it infrastructure for the internet of things: How can nfv & sdn help for facing the upcoming challenges," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015, pp. 64–69.

[21] S. Bagchi, "Unlocking the potential of sdn and nfv." [Online]. Available: http://blogs.windriver.com/wind_river_blog/2016/08/unlocking-the-potential-of-sdn-and-nfv.html

[22] C. Mathas, "What's next for telco sdn/nfv? fast growth, slower deployment." [Online]. Available: https://itu4u.wordpress.com/2016/06/28/whats-next-for-telco-sdnnfv-fast-growth-slower-deployment/

[23] F. A. Khan *et al.*, "Virtualized epc: Unleashing the potential of nfv and sdn," in *25th European Regional ITS Conference, Brussels 2014*, no. 101426. International Telecommunications Society (ITS), 2014.

[24] M. Dano, "How verizon, at&t, t-mobile, sprint and more stacked up in q2 2016: The top 7 carriers." [Online]. Available: http://www.fiercewireless.com/wireless/how-verizon-at-t-t-mobile-sprint-and-more-stacked-up-q2-2016-top-7-carriers

[25] J. Doherty, *SDN and NFV Simplified: A Visual Guide to Understanding Software Defined Networks and Network Function Virtualization*. Addison-Wesley Professional, 2016.

[26] A. D. Little, *Reshaping the future with NFV and SDN*. Bell Labs Alcatel-Lucent, May 2015.

Chapter 9

Challenges in 5G: Social, Technological, and Economical Perspectives

Jérôme Oesch, Christian Schneider, Yannic Blattmann

As the rollout of 4G communications networks has taken place in most parts of the world, research has started on new technologies that will pave the way to 5G communication networks. While no technologies have been standardized and adopted by key standardization bodies like IEEE, developments in all directions of wireless networking are open. Many articles have been published, foreseeing the upcoming 5G communications network technologies. Hence, in this paper we aim to summarize key challenges of 5G mobile communications and reveal key technologies that have been suggested by academia as well as representatives of the industry, and implications on economical and social perspectives. First, key technologies, that will most likely play a crucial part in upcoming 5G mobile communications are reviewed. These are grouped in topics as the evolution of radio access technologies (RATs), composite wireless infrastructures and heterogeneous network deployments as well as traffic offloading and hybrid topology networking, and the introduction of intelligence (reducing hardware limitations with software approaches). Along with those key topics, we discuss economic perspectives of the technologies and how they will impact cost and quality of service (QoS). Hereafter, we present social perspectives on how these new technologies will impact the user's life, if there are noticeable changes and how the user's quality of experience (QoE) will be affected. Our findings include the following: For 5G, we may expect improvements in bandwidth capacity, increased data rates and reliability as well as better energy efficiency. Concerns may arise in areas of environment protection and data security.

Contents

9.1 Introduction

Since the introduction of 4G, many new technologies and ideas have evolved on how a future 5G radio access network may look like. Today, the 4G network is mainly based on macrocells that use the LTE standard, together with older 3G and 2G cell standards such as HSDPA, UMTS, EDGE or GPRS. Network deployments with big macrocells spanning over big areas has proven to be sufficient for recent years, but with the upcoming Internet of Things and a general rise in user devices in the network and higher bandwidth usage, this architecture will sooner or later hit the wall of its performance. This can be observed in congested areas, where cellular network providers try to offload traffic from macrocells to Wi-Fi hotspots with auto-login.
While these technologies are discussed themselves in a broad manner, economical and social consequences are often left aside. Economical aspects do have an influence on their acceptance in real life. In addition, the overall goal of such new technology is to provide higher quality of service (QoS) as well as improving on the quality of experience (QoE).
The goal of this paper is therefore to outline new technologies and also to cover economical and social consequences such technologies could have, should they be rolled out. In the first part of this paper, possible new technologies are listed, which have been proposed and have been approved by a major part of the researching community.
Demestichas et al. [15] give an overview of upcoming, possible technologies that could be used in 5G (see Figure 9.10). There are three main directions that current thoughts in technology for 5G are heading. First, there is the evolution in RATs themselves, where new technologies are enhancing existing technologies, as well as a more flexible spectrum management, which is discussed in Section 9.2.1. Second, there is the idea of shrinking cells to better fit their environment, as well as traffic offloading strategies to free macrocells, which is discussed in Section 9.2.2. Third, there is the introduction of intelligence, where backbone architecture is revised to give more performance to existing hardware. This includes concepts of outsourcing hardware control to the cloud, or to implement intelligent software that is able to solve frequency interference, which is discussed in Section 9.2.3.
In the second part of the paper, economical perspective of the new technologies is discussed, where we focus mainly on QoS and costs. In the third part, social perspectives are discussed, where focus is set on QoE, cost and ecology. In the final conclusion, technical, economical as well as social perspectives are evaluated and compared, and an outlook to the upcoming 5G network is given.

9.2 Technological Perspective

Currently scientists discuss new technologies that will most probably be applied in and standardized for 5G. In this section, such technologies are presented, their benefits and some of their challenges.

9.2.1 Evolution of Radio Access Technologies

Under Radio Access Technology (RAT) we understand the underlying physical connection method for a radio based communication network. This section describes the changes in Radio Access Technologies (RATs) which are expected for 5G. The following two Sections (9.2.2 and 9.2.3) will then focus on the application of these technologies and further discuss how limitations of hardware can be bridged with software approaches.

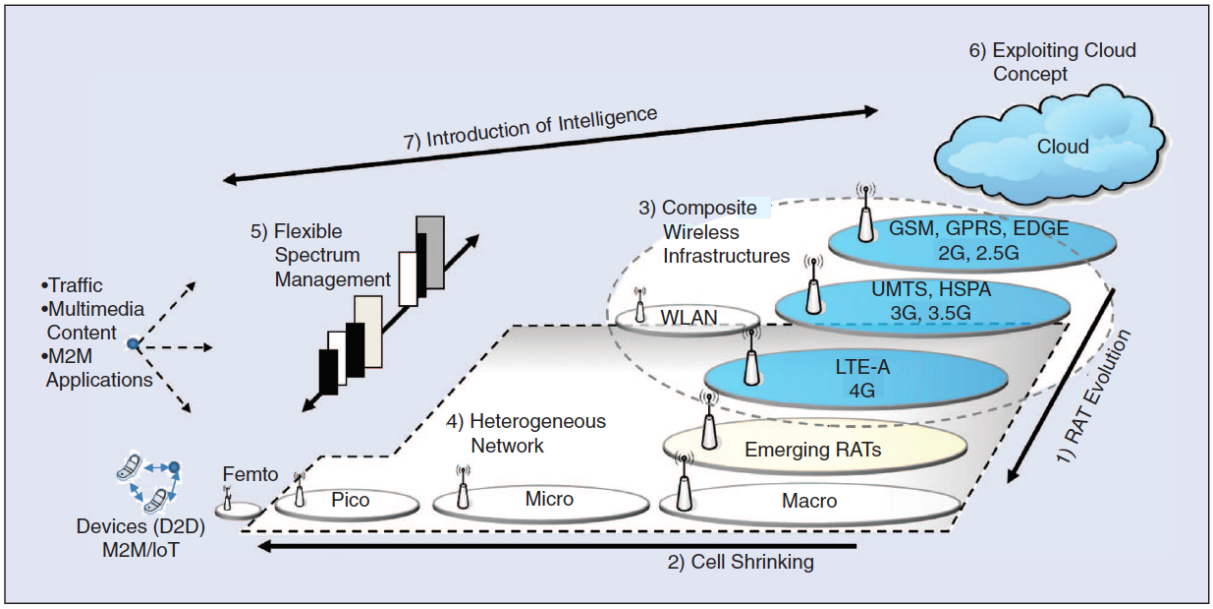


Figure 9.1: Overview of upcoming 5G technologies [15]

9.2.1.1 Asymmetric traffic & full duplex radio

In a full duplex communication in a single band, both links (uplink and downlink) have a symmetric link capacity [24]. However, uplink and downlink may not have to send equally much data, which means that either uplink or downlink may be underutilized, which is not efficient. Malik et al (2015) [24] suggest a technique aiming for maximizing the downlink data rate . This is called asymmetric traffic.

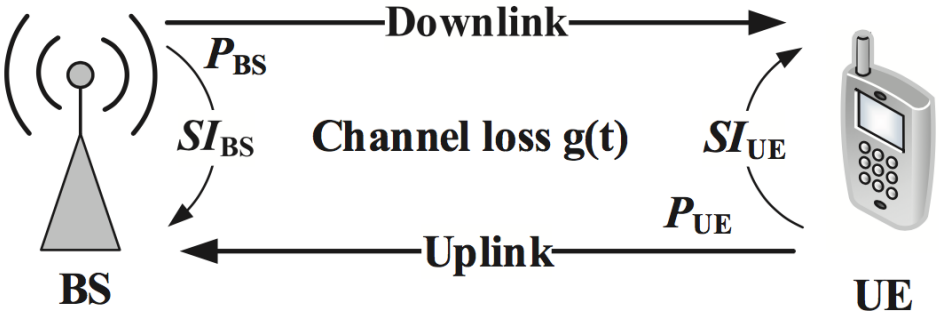


Figure 9.2: Full-duplex link [24]

The authors' suggestion is to increase power in order to be able to transmit more bits per symbol. This enables the user to take advantage of high download performance.

9.2.1.2 Millimeter wave (mmWave)

In order to increase data rates in 5G, millimeter wave communication (mmWave) is considered. mmWave broadens the available wireless frequency spectrum (see figure 9.3). According to Ma et al (2015), today there is a bandwidth shortage in today's frequencies under 3 GHz [23]. A number of researchers and companies in the field are researching on making mmWave feasible. mmWave will enable to increase the available bandwidth and, thus, data rates and capacity. Two of the challenges, taken from the mentioned authors, are the following: On the one hand, "antenna arrays are considered as a key technique to achieve mmWave communications." However, the narrow beams raise environmental concerns. On the other hand, some of the involved technical components, analog-to-digital

converters and digital-to-analog converters, have high power consumption and are, thus, not efficient.

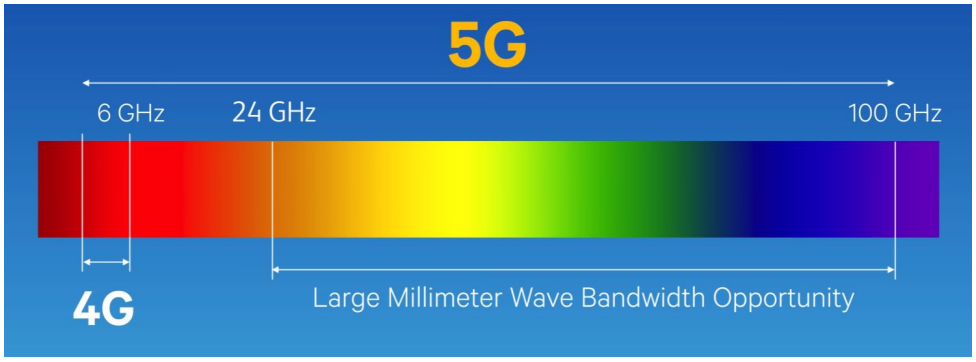


Figure 9.3: Additional spectrum in mmWave [31]

9.2.1.3 Massive MIMO

MIMO (multiple in, multiple out) allows a mobile user to access data, communicating with a sender over several antennas simultaneously instead of just one (see figure 9.4). This chapter is based on information from Larsson et al (2014) and Quoc Ngo (2015) [19, 26]. An improvement to MIMO was intended with Multi-User MIMO itself, which uses the same amount of antennas, but strives for efficiency gains with frequency-division duplex operation. However, Multi-User MIMO is not a scalable technology. Massive MIMO, being a form of Multi-User MIMO where the number of antennas is increased, strives for improvements in both scalability and performance by increasing the number of antennas, extensively using low-power components. Time-division multiplex operation reduces channel-estimation overhead, which occurs with Frequency-division multiplex. Benefits of Massive MIMO include significant improvements in throughput and radiated energy efficiency and it is more robust against intentional jamming. Challenges include the following: As Massive MIMO is different from current practices, respective low-cost components which work together well have to be found. Energy consumption of the respective hardware has to be reduced.

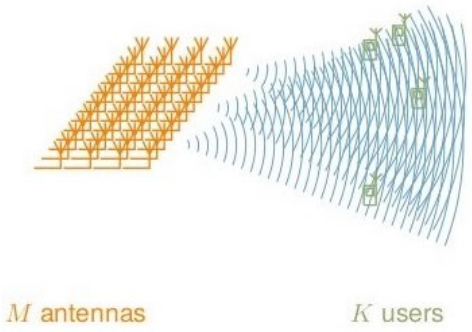


Figure 9.4: Antennas and users in Massive MIMO [5]

With Massive MIMO, all the complexity is at the base station. Massive MIMO is scalable as Base Stations learn about channels in uplink training and the number of Base Station antennas can be increased without increasing the channel estimation overhead.

9.2.1.4 Ultra Low Latency

In 5G, data rates are expected to be increased and latency decreased. Thus, Ultra Low Latency is a key concept in 5G. This is especially beneficial for games and augmented reality [25]. The author also states that end-to-end delays under 1 ms are relevant for

M2M communications and social networks. The author suggests using sub-networks to manage a proportional part of the traffic load and decrease the round trip time. The number of hubs can be reduced to make data travel faster.

9.2.1.5 Cognitive Radio

A common challenge in wireless communication is the fact that the wireless spectrum is scarce. At the same time, the capacity of the spectrum is poorly used [23] and users require an increasing amount of data. However, available channels are not used uniformly, meaning they do not use them to same extent all the time [13]. Besides the primary users of channels (users who are licensed for a channel), secondary users (who are licensed for a different channel) can join in the usage whenever the spectrum is not fully used by a primary user. To enable multiple cellular operators to share their radio spectrum, mechanisms for managing possible interferences are being developed. Cognitive Radio strives for increasing primary user satisfaction besides letting secondary users join the user services [13]. Cognitive Radio consists of following four major functionalities [23, 6]:

- **Spectrum management:**
  Capturing the best available spectrum to meet user communication requirements.
- **Spectrum sensing:**
  Detecting unused spectrum and sharing the spectrum without harmful interference with other users.
- **Spectrum mobility:**
  Maintaining QoS during the transition to other spectrum.
- **Spectrum sharing:**
  Providing the fair spectrum scheduling method among coexisting network users.

In what follows, the latter three functionalities are explained in further detail.

Spectrum Sensing Spectrum sensing aims to detect unused spectrum and sharing the spectrum without harmful interference with other users [6]. Akyildiz et al (2006) give following overview of available detection techniques [6].

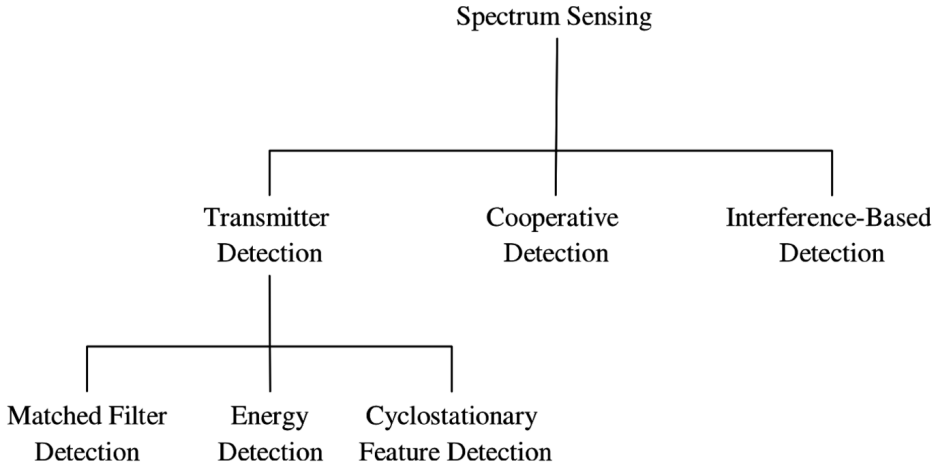


Figure 9.5: Classification of spectrum sensing techniques [6]

With **Transmitter Detection**, secondary users aim to detect whether primary users are using a certain portion of the spectrum or not. There are three major kinds of Transmitter Detection techniques: With **Matched Filter Detection**, secondary users look for specific

patterns which are sent over a channel like preambles or spreading codes in order to assess whether a channel is already in use. With **Energy Detection**, secondary users measure the the energy of received signals to assess whether a channel is already in use. With **Cyclostationary feature detection**, secondary users couple used signals with waves which exhibit certain patterns (cyclostationary patterns).
Cooperative Detection is, in principle, more accurate than Transmitter Detection. The uncertainty in a single user's detection can be minimized. This is enabled by a central controller creating an occupancy map for the network. The disadvantage here is that the exchange of occupancy map data creates additional traffic.
A third technique for Transmitter Detection is **Interference-Based Detection**. It allows a secondary user to assess whether a channel is already in use by sensing the level of interference called interference temperature. If the interference is sufficiently low, a secondary user can use the channel. It is challenging however to effectively measure the interference temperature, performing this task in multi-user networks and improving the performance in detecting.

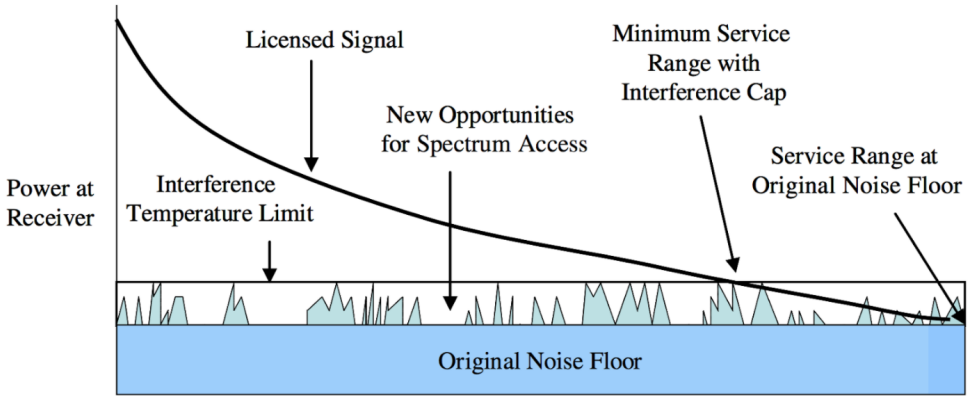


Figure 9.6: Interference temperature model [6]

Spectrum Sharing Spectrum sharing means to provide a fair scheduling method among coexisting users [6] – allowing secondary users to use primary user's spectrum whenever the latter ones do not use it. Today sharing mechanisms are greedy and competitive. Rebato et al (2016) argue that "sharing spectrum licenses increases the per-user rate when antennas have narrow beams, and that if network operators share their licenses, they can achieve the same per-user median rate as if each had an exclusive license with more bandwidth." [27].
Akyildiz et al (2006) suggest two possible levels of sharing sharing spectrum as shown in the figure above [6]. **Inter-Network Spectrum Sharing** enables multiple antennas to share spectrum. **Intra-Network Spectrum Sharing** enables secondary users to access spectrum over primary users.

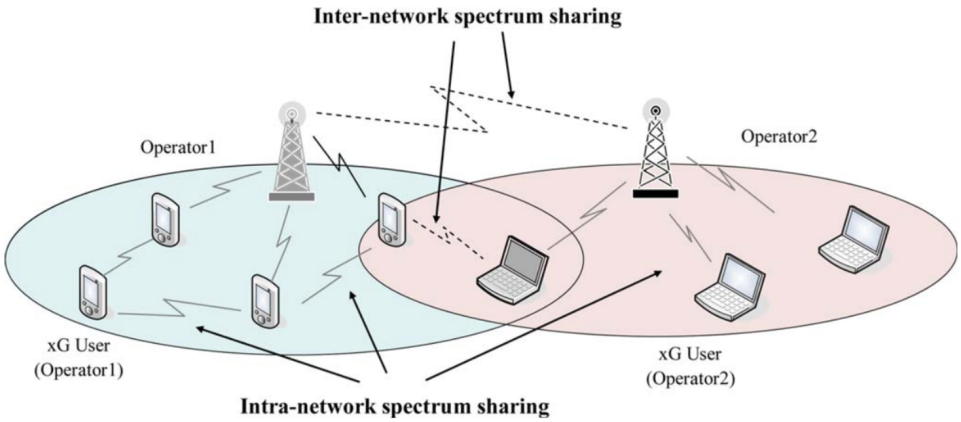


Figure 9.7: Inter-network and intra-network spectrum sharing [6]

Furthermore, Akyildiz et al 2006 present the major steps which enable Spectrum Sharing for mobile users [6].

1. **Spectrum sensing**
   Detecting unused spectrum (as previously discussed).

2. **Spectrum allocation**
   Allocating spectrum to users.

3. **Spectrum access**
   Users accessing the spectrum.

4. **Transmitter-receiver handshake**
   Establishing the connection between transmitter and receiving user.

5. **Spectrum mobility**
   Secondary users need to leave their part of the spectrum when primary users need it. Then, secondary users have to be assigned to new portions of spectrum (as previously discussed).

According to Akyildiz et al (2006) following distinction of spectrum sharing techniques can be made [6]: In terms of architecture, spectrum sharing can be controlled by a central entity (centralized) or based on local policies (distributed). Spectrum allocation can be handled either in a cooperative manner or in a non-cooperative manner. Non-cooperative spectrum allocation is computationally easy, because a user can distinguish himself/herself whether a portion of spectrum is used or not. Cooperative spectrum allocation is better in overall efficiency, because allocation decisions are taken in accordance with multiple users, however, this makes it computationally more complex. In terms of spectrum access technique, a distinction between overlay access technique and underlay access technique can be made: While with (THE) overlay access technique(S) a user accesses the spectrum over an unused portion of the spectrum in order to find out which part of the spectrum can be used, with underlay access technique a user produces noise while using a certain part of the spectrum to show other users that their part of the spectrum can not be used in that specific moment.

Spectrum Mobility
When a primary user needs the portion of spectrum which is in use by a secondary user, the latter will then have to be assigned new portions of spectrum. In this situation, Spectrum Mobility enables to maintain seamless communication [6]. The same challenge applies, for instance, when a user changes location. The *spectrum handoff* which is required for this shift of location takes place through an algorithm which has to decide which other

portion of spectrum is most promising in terms of performance [23]. According to Akyildiz et al (2006), Spectrum Sensing can provide necessary data for the algorithm to make that decision. The author also outlines following two research challenges in terms of spectrum handoff among other challenges in the field: How can delay in the spectrum handoff process be reduced? How can running applications be transferred to a different frequency band without letting the applications suffer too much from performance degradation?

9.2.1.6 Dynamic Spectrum Access (DSA)

Dynamic Spectrum Access (DSA) enables dynamic sharing of spectrum bands [15]. Two of the most common DSA models are Authorized-shared access (ASA) and Licensed-shared access (LSA). Following information is taken from Lehr et al (2014) [20]. "Most ASA/LSA systems propose a centralized database mechanism to control spectrum allocations." A user receives information about available communication paths. While an assigned portion of spectrum is used, the user has access rights to the respective portion of spectrum (see figure 9.8). The access right is protected by interference. ASA and LSA allow predictable quality of service for licensed and unlicensed users and is, thus, suitable and reliable for applications like phone calls. So far, it is not envisaged that end-user devices need sensing capabilities for ASA and LSA.

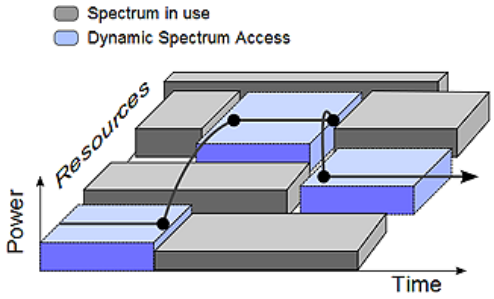


Figure 9.8: Dynamic Spectrum Access (DSA) [32]

9.2.1.7 NOMA & SCMA

Non-orthogonal multiple access (NOMA) and sparse code multiple access (SCMA) are two rather *new wireless multiple access techniques*. Examples for other wireless multiple access techniques are time division multiple access (TDMA) or code division multiple access (CDMA). Following information on NOMA and SCMA is taken from Ma et al (2015) [23].
NOMA multiplexes users in a power domain. Multiple users are allocated at different power levels depending on their channel condition. Here, the author presents an example with two users, user 1 being close to a base station and user 2 close to a boundary of a cell. Using NOMA, user 2 is served with more transmission power as its channel condition is worse than the one of user 1. Using different transmission power, both user 1 and user 2 can be served simultaneously. NOMA exhibits better spectral efficiency than existing multiple access techniques like TDMA or CDMA, because it takes channel conditions into account. SCMA exploits sparse codes to multiplex users. It is codebook-based and non-orthogonal. An encoder has to be in place here. Benefits of SCMA includes reduced complexity and robust link adaptation.

9.2.1.8 Energy performance in 5G

In Figure 9.9, Taha (2012) shows the increased global mobile traffic [29]. This increase also impacts energy consumption and, thus, the environment.

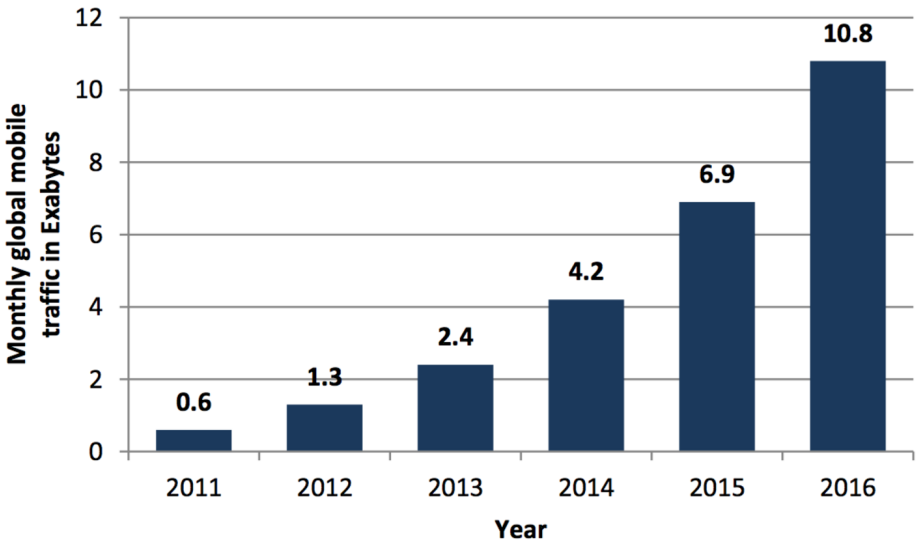


Figure 9.9: Global mobile traffic per month [29]

Furthermore, as explained in the chapters before, some of the components which are expected to be used in 5G consume a relatively high amount of energy. According to Ericsson (2015), energy efficiency helps reducing operational cost and total cost of ownership for users, and makes 5G more sustainable [16]. The company suggests principles which are intended to support energy efficiency in 5G:

- **"Always Available" instead of "Always On"**
  Developing designs for processor sleep can help reducing energy consumption of devices which, otherwise, would be running constantly and consuming energy.
- **"Only be active and transmit when needed"**
  Ultra-Lean Design enables devices to have more time without transmission and allowing more time in sleep mode. This reduces energy consumption.
- **"Only be active and transmit where needed"**
  For instance, through separating user data and system-control plane functionality, more data processing happens on end-user devices and less data has to be transmitted. This can reduce energy consumption.

9.2.2 Blended Infrastructures & Traffic Offloading

In today's cellular network infrastructure, so called macrocells are dominating the network design, where single antenna towers with often multiple antennas provide a cellular network with a often very large cell radius. In addition, some cellular operators have started non-overlay traffic offloading via Wi-Fi hotspots that are used in highly congested areas where a lot of traffic occurs also often called "Composite Wireless Infrastructures" [10, 15], (see Section 9.2.2.2). These base stations have to be differentiated from cellular network base stations, as they do mostly not supply cellular access and are connected to the internet directly, another example being UDRANETs. Heterogeneous network deployments on the other hand, base on one RAT standard only, for example 4G / LTE.

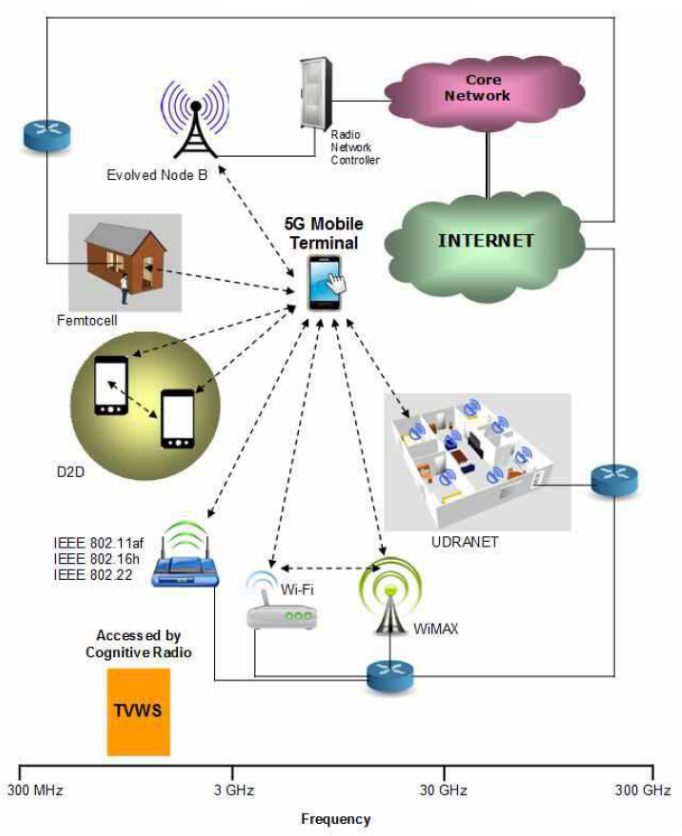


Figure 9.10: Overview of the upcomming blended infrastructure in 5G [10].

Even though this design mostly satisfies our today's needs, it is often already at the limits of maximum performance. This is due to overloaded macrocells in highly congested areas, or due to bad reception inside new, better isolated buildings. Coming up Internet of Things traffic will strain the already stressed network additionally. In many of these cases, due to physical laws of frequency band interference or regulations based on maximum radiation thresholds, building new antennas does not solve the problem.
It therefore can be foreseen that upcoming 5G technologies will depend on a variety of new, different technologies to better address the needs for capacity, coverage and high quality of service instead of relying on a single technology, topology or regulatory approach [2]. In detail, 5G will combine the use of cellular, local and personal area networks, as well as short range device technologies and topologies, and will affect current regulations concerning frequency management. Chavez et al. [10] mention two main reasons for this theoretical development. Firstly, that there most likely will not be a single, standalone technology available that will be able to provide a sufficient capacity improvement to accommodate all market sectors, and secondly that the return on investment is much higher on infrastructure that can be utilized to its maximum utility.
In the following subsections, the key technologies envisioned to be part of the future blended infrastructure are outlined. These technologies contain heterogeneous network deployments (see section 9.2.2.1), composite wireless infrastructues (see section 9.2.2.2), hybrid topology networking (see section 9.2.2.3) and caching (see section 9.2.2.4).

9.2.2.1 Heterogeneous Network Deployments

A technology most likely going to be part of 5G that will be able to tackle the problem of rising demand in bandwidth is heterogeneous network deployment, where load on macrocells is reduced by deploying low-power cellular base stations in indoor and outdoor areas. The base stations are directly connected to the conventional cellular network backbone, and are called femto-, pico-, nano- and microcells depending on the range of the indi-

vidual base station and their purpose (or called "Local IMT small cells" [11], depending on reference). In comparison to today's practice of mostly WiFi-hotspots, all mentioned cell types are able to handle voice and data traffic with QoS assured [10]. While pico- to microcells are usually operated and frequency-wise configured by the cellular network provider, femtocells can act on their own and find the best available frequency to operate on and the radiated power output themselves. They are often provided by cellular network providers to individuals for home use already today, and are called cognitive femtocells [10] due to their capabilities. This usage is not unproblematic though, as these femtocells currently use the same frequency ranges as macrocells and therefore can interfere with each other. However, this problem is reduced by a much lower transmission power of small cells, which provides a more flexible spectrum management and also allows power savings. Other differences between the cell types can be found in supported moving speeds of connected users, where small-radius cells only support nomadic speeds and big cells, such as current macrocells, support much higher speeds of up to 350 km/h (Chen, S). In addition, indoor propagation characteristics differ a lot from outdoor propagation, which results in higher possible performance benefits to small cells.

To solve the problem of interference between base stations, one major challenge of this new technology is to allocate frequency spectrum for these small cells. Opinions differ whether different frequency spectrums (spectrum split) should be used for the different sized cells or not (spectrum sharing, see also section 9.2.1.5). Haddad et al. [18] discuss in their paper that frequency sharing leads to the problem of co-channel interference, and that this can already be observed in cases where telecommunication providers provide femtocells to private users to improve cell-phone reception in their homes. One possible solution allowing the usage of shared frequency bands between femto- and macrocells is to share the band in a time division multiple access manner (TDMA), where femtocells as well as macrocells have predefined timeslots in which they are allowed to send data or not. Another approach is Software Defined Radio 9.2.3.1.

On the other hand, splitting the frequency spectrum would solve the problem of interference, but as telecommunication providers do not own additional frequency bands to allocate to the femtocells, the already too narrow spectrum for macrocells would have to be constrained additionally. Chen et al. [11] suggest, that instead of using already allocated spectrum, to use ultra-wide bandwidth in very high frequency spectra ($6 \sim 60$ GHz) as dedicated femtocell spectrum. Both of these suggestions however do not solve the problem of interference between the femtocell base stations themselves, where frequency sharing has to be addressed as well. This has less influence on base stations provided by telecommunication providers at known locations, as in this case spectrum and transmission power can be managed between the base stations. The problem mostly emerges if femtocell base stations are supplied to private users to improve reception in their homes, as cognitive femtocells, as location and neighborhood of these devices is unknown. With cognitive radio, this problem can be tackled however (see section 9.2.1.5).

Another approach to increase throughput, the so called Flexible Separation and Coordination of C/U Planes, is mentioned by Chen et al. [11], where multiple local small cells are connected to the same logical control unit as its spanning macro cell. This decoupling from the user plane and the control plane has advantages, and no noticeable disadvantages for the user. While the macrocell is sending in its own frequency band, the small cells send in other frequencies, and a device can be connected to multiple cells at the same time. This allows traffic offloading of macrocells, without the expensive changeover from one cell to another one. Advantages are much higher local transfer speeds, less transmission power needed, dynamic frequency usage can be supported and QoS rises. Challenges for this approach are the need of a dual receiver of the user.

9.2.2.2 Composite Wireless Infrastructures

Composite wireless infrastructures depict approaches on the interworking of cellular systems depending on different RATs. In contrast to heterogeneous network deployments, composite wireless infrastructures use different radio access technologies, which is targeted at the improvement of application provisioning [15] .
In the example of Wi-Fi, if a device is in vicinity of a Wi-Fi hotspot, data traffic is routed via the hotspot instead of the macro cell the device is primarily connected to. For cellular operators, this has the big advantage of using free, unlicensed spectrum instead of expensive and congested frequency bands [9].
Current challenges in further development of Wi-Fi are to increase its spectral efficiency to allow more users on a Wi-Fi network, and that Wi-Fi currently does not provide QoS differentiation due to its MAC protocol [10]. Therefore, voice services are in most cases still delivered via the core network. There however exist implementations of on-the-spot traffic offloading, where Wi-Fi is prioritized over cellular connections, and as soon as the user leaves the Wi-Fi's range, unfinished data transfer is finished over the cellular network. This also allows for VoIP, as the phone can switch to cellular infrastructure when leaving the Wi-Fi's range. Another approach exists where Wi-Fi networks have overlapping channels, as a recent study has shown that Voice over IP (VoIP) signals are not significantly degraded in the 2.4 GHz band by the resulting channel interference [9].
For full offloading of VoIP signals, Chàvez et al. [9]. suggests using standards defined IEEE 802.11, one of which standards is called White-Fi. White-Fi uses the free TV white space spectrum through cognitive radio for a transmission with conventional Wi-Fi technology. Another key technology, WiMax, also called "Wi-Fi on steroids" [33] , could also be an enabler for 5G. It uses a similar architecture as Wi-Fi, but can be used in a much bigger radius than Wi-Fi. Chàvez et al. [9] see potential in using WiMax in Wi-Fi backbones, but mentions that this technology has not been considered to interoperate with cellular networks, which would need some additional standardization. It has to be mentioned though that IEEE voted to resolve group 802.16 containing WiMax. This will most probably amount into a death sentence, reason being its unsuccessfulness on the market.
According to Chàvez et al. [9], another of the concepts which is expected to be applied in 5G is Ultra-Dense Radio Access Networks (UDRANETs). UDRANETs are expected to consist of low-power access nodes allowing high traffic capacity over short-range links, arranged few meters apart of each other. Probably UDRANETs will be run in a frequency range between 10 GHz and 100 GHz. So far, this range has not been relevant to the industry meaning that new transmission and access technologies have to be developed. These new transmission and access technologies will require spacely developed millimeter waves (mmW) though (see section 9.2.1.2).
UDRANETs are envisioned to be used to offload traffic for extremely high data rate applications, with a major challenge being to produce low-cost mobile terminals able to operate in super-high frequency and extremely-high frequency bands. Thanks to the usage of mmW, it will be possible to pack more antennas into terminals, allowing the implementation of massive multiple-input multiple-output (M-MIMO, see section 9.2.1.3) for an improved throughput and a lower latency.

9.2.2.3 Hybrid Topology Networking

Relay Technologies in form of repeaters are a feature proposed by Chen et al. [11] and Ma et al. [23]. For this feature, the cellular network cells can have relays that improve its topology. With this mesh-grid, machine to machine communication can be established very easily, without a need for a stationary core network. A user mode will enable single

devices to be a node of the network. Relays can be divided into in-band relay and out-band relay. In-band relay is sharing the frequency space, whereas out-band relay utilizes dedicated frequency ranges. Both support time division duplex (TDD) and frequency division duplex (FDD). Chen et al. [11] believe that TDD has more advantages over FDD in this specific scenario, as one frequency band suffices for up- and downlink when compared to FDD, where for up- and downlink two frequency bands are needed.
Relaying has been specified in LTE-Advanced (LTE-A) release 10, and is named Relay Node. Ma et al. [23] mention the ability of relaying wireless communication systems to send a certain message through various routes to the receiver. This would be employed large-scaled and would improve coverage and throughput of our existing network infrastructure. Underlying technologies would include relay selection, relay combining and distributed space-time coding [23]. Relay selection handles the best choice of relay points to a receiver opportunistically. Relay combining can be used in a system where receivers have full knowledge of channel state information. Relays can be then combined and offer the better performance. A distributed beamforming coefficient is multiplied at each relay, respectively. Distributed space-time coding is targeted at making the relaying system more reliable. To allow for spatial diversity (data is sent over multiple nodes to guarantee complete retrieval at the receivers side) in relay systems, the nodes need to be time-wise synchronized. Distributed space-time coding is a work proposed to solve this issue in an asynchronous manner.
Device to device communications (D2D) are related to relay technology. Both will use similar technologies for maximum throughput. Direct device to device communications have however less current use cases - up until now. The internet of things or caching (see section 9.2.2.4) for example could play a key role to offload traffic from the core network. Chàvez et al. [10] mentions the similarities to ad-hoc networks, with the key difference that they use unlicensed spectrum instead of licensed.
Chen et al. [11] differs D2D based on frequencies used: co-channel frequency and dedicated frequency D2D. In co-channel frequency D2D, the user device operates in the same frequency band as the one between cell and user. Advantages are the reusability of the current receivers and transmitters, but on the other hand serious interference to the cellular system due to the mobility of user equipment. In addition, transmission rates will be quite low compared to dedicated frequency D2D. Dedicated frequency D2D on the other hand will own its own frequency range, most possibly somewhere in the higher frequency spectrum, which also reduces interference greatly. This allows for higher transfer speeds, but requires for new receivers and transmitters in the devices, plus two sets of them for communication between cell and the user, as well as the user with another user. Chen et al. [11] predict that the latter solution will be the main type of D2D, while co-channel frequency D2D is also kept to support D2D between legacy UE. Pierucci [25] and Demestichas et al. [15] make clear, that this system needs cognitive, intelligent management mechanisms to control and manage all these networked devices. And Pierucci [25] also mentions the battery lifetime of single nodes, in our case user devices, as although the device is not used by the user himself, its battery life will decrease.

9.2.2.4 Caching

In order to be anticipatory, we have to assume that the data patterns from mobile users are, to a certain degree, predictive. With such a prediction the peak load can be flattened by preloading the data to the user's device, in off peak hours, before the user requests that data. And if the request is initiated the data is pulled from the devices cache instead of the wireless network and thus doesn't consume traffic in peak hours. Over the last years, predictive analytics and big data have made significant progress using machine

learning techniques to collect and analyze huge amounts of infrastructure logs to produce predictive information for outage prediction and content recommendation [21].
For this matter new machine learning tools should be developed to minimize the data cached on the users device and data not preloaded to the users device but requested. Analyzing user traffic and taking the user's social networks into account can enhance such tools. But you might not get all these information from every user and thus need to aggregate a statistical solution to predict his/her data pattern from other users and their social network. Caching content locally at the small base stations and the user terminals can reduce the backhaul significantly, especially if the network is flooded with similar requests [7].
According to Bastug et al (2014) backhauling is of utmost importance before a roll-out of Small Cell Networks (SCNs) [7]. In this network model, the Small Base Station (SBS) are deployed with high capacity storage units but limited backhaul links. To optimise the use of the storage the authors improve the proactive caching procedure from Bastug et al (2013) where the most popular data were stored until the storage was full, by caching procedure with a training and placement part [8]. In the training part the goal is to estimate the population matrix by the users preferences for every SBS's own model. In the placement phase the storage is quickly filled based on the popularity in the SBS's model. In Bastug et al (2014), the findings show that the proactive outperforms the reactive caching [7].

9.2.3 Introduction of Intelligence

In what follows, software approaches which reduce hardware limitations will be explained.

9.2.3.1 Software defined radio (SDR)

A software defined radio (SDR) is a "radio in which some or all of the physical layer functions are software defined" [3]. In this context, a radio can be any kind of device that wirelessly transmits or receives signals in the radio frequency (RF) spectrum, e.g. phone, personal computers, televisions [3].
The Wireless Innovation Forum, it was called SDR Forum before, distinguishes between five tiers of SDRs, where Tier 0 is the hardware radio, Tier 1 is a software controlled radio, Tier 2 the software defined radio, with software modulation, bandwidth, frequency range and frequency bands controlled. Tier 3 fully programmable radio and Tier 4 the ultimate radio with receiving satellite transmissions in real time, but the last two exists until now only on paper. SDR can be the solution to the saturation of frequency bands with it's self adaptive mechanisms [14].
SDR provides the flexibility to change the spectrum bands through programmable radio frequency (RF) frontend. RF frontend refers to the hardware from the antenna to the frequency mixer (see figure 9.11). The SDR transceiver can act in real time on the radio environment and change transmitting parameters, and devices in the network can have multiple radios with an SDR transceiver. These devices may also transmit or receive different frequency bands, which enhances the flexibility and spectrum allocation. Therefore SDR enables wireless networks with efficient spectrum assignment through its extensive use. This enhances the spectrum utilization for 5G [22].

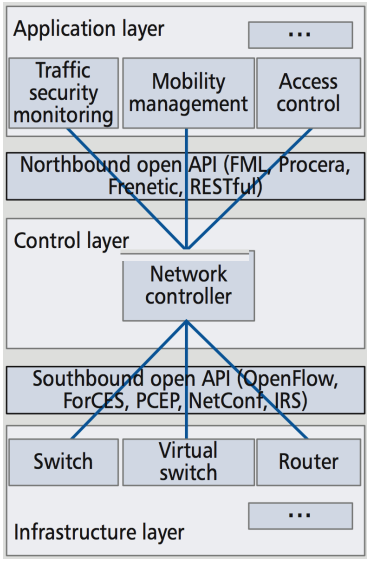


Figure 9.12: SDN Layering concept [28]

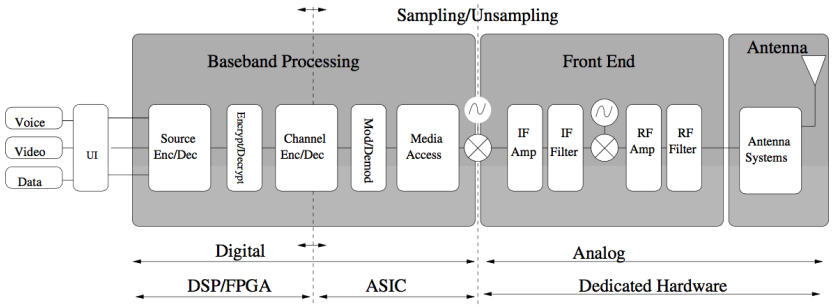


Figure 9.11: Analog and digital part of SDR [14]

Cognitive radio (see 9.2.1.5) is a development of SDR. In SDR, a user device scans for free frequencies and has multiple connection technologies in one single interface. This means one device can receive signals from WiMAX, Wifi, GSM, or LTE [12].

9.2.3.2 Software defined networking (SDN)

Software defined networking (SDN) is a software approach to create a flexible service to reduce hardware limitations. SDN is a new emerging network architecture, where the network controller is directly programmable. The centralized controller is responsible for allocating traffic to network elements in the network layers as shown in the Figure below. The network controller maintains the intelligence and state of the entire network and can, for this reason, chose the best routing flow control rules to the heterogeneous network (see 9.2.2.1) devices from various manufacturers. Furthermore, network devices themselves do not have to understand and process all the protocol standards, but just accept and follow the instructions provided by the SDN. The network controller can present a global and united view and resources to the upper layer network applications, as a single logical switch. This allows to easily create, test and deploy new network applications in a short amount of time [15, 17, 10].

With the northbound open APIs, between SDN controller and application layers, applications run on an abstraction of the network, which optimizes network services and capabilities without regarding the details of the implementation. As a result, computation, storage, and network resources can be optimized [28]. For the southbound, the dominant protocol is OpenFlow, which is the first standard communications interface defined between the control and forwarding layers of an SDN architecture [17]. As mentioned in the ONF White Paper [17], OpenFlow allows direct access and modification of

the forwarding plane of network devices, both physical and virtual. And this is required to move the network control to a logically centralized control software.
OpenFlow has to be installed on both sides of the interface between network infrastructure devices and the SDN control software. OpenFlow uses pre-defined match rules, which can be statistically or dynamically programmed by the SDN control software, to identify the flow of network traffic. It also allows to define how the traffic should be guided through network devices, through parameters as usage patterns, applications, and cloud resources. Furthermore, it allows the network to be oriented on a per-flow basis and act in real-time to changes at the application, user, and session levels, which current IP-based networks do not provide. As all flows between endpoints must take the same path in the network, it does not matter if they have different requirements. Since OpenFlow SDNs are well adaptable, it can easily be deployed in existing networks, both physical and virtual. And network devices can support both OpenFlow and traditional forwarding. This makes it easy to seamlessly integrate SDN architecture into an existing infrastructure and provide a migration path for segments of the network, which have the most benefits or/and need of SDN [17].
According to Chavez et al [10], disadvantages of SDN include security aspects, because in case of an attack on the controller huge areas of the network would be compromised. And since the controller has a global overview over the network status information could be accessed. Then again, the network security is improved according to through the centralized and automated management of uniform policy enforcement [17].

9.2.3.3 Network function virtualization (NFV)

As discussed by Demestichas et al [15], NFV is an additional technology to SDN, on which SDN can be implemented. NFV is a new way to configure an end-to-end network infrastructure with virtualization technology. The network function of a device is implemented in a software package, which is running in virtual machines created in generic high-volume hardware servers and storage devices. This virtual infrastructure is connected by high-volume network switches and organized by the orchestration (see figure 9.13). The software functions for the devices are separated from generic hardware. And automated orchestration will automate the installation and management of the virtualized network functions on the generic hardware. This would make testing new network functions by simply installing or upgrading the software package, which is run by the servers, much easier [15, 28].

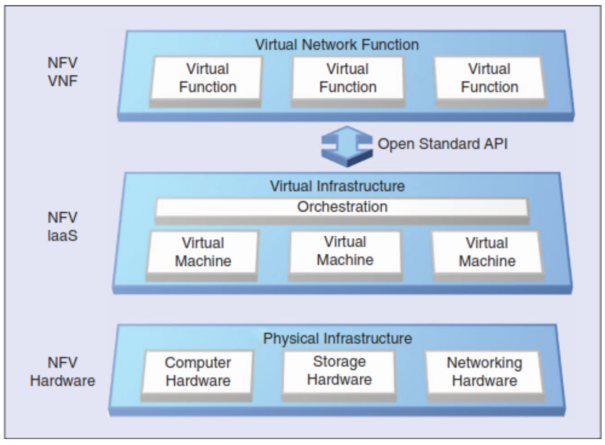


Figure 9.13: NFV concept overview [15]

9.2.3.4 Fusion Net

As described in Demestichas et al. [15], Fusion Net or multilayer and multistream aggregation (MSA) technology is where a host layer, which consists mainly of macrocells, provide the omnipresent and reliable basic user experience . The additional boosting layer consists of many infrastructure elements (e.g. boosting port from small cells), spectrum (e.g. boosting carrier), and multitechnologies (e.g. boosting RAT). Through the three boosting sublayers the best possible performance can be achieved, all streams from the sublayers are aggregated by MSA technology with the host layer (see figure 9.14). To minimize handoff and interference the user device is assigned to a macro cell (host layer), while the traffic is dynamically transmitted and aggregated by a boosting layer. The decision-making entity is located centrally in the radio access network (single radio controller/Cloud RAN), where the air interface and the backhaul and computing resources are located and intelligent algorithms can be implemented to make the RAN transparent for the user and the core network [15].

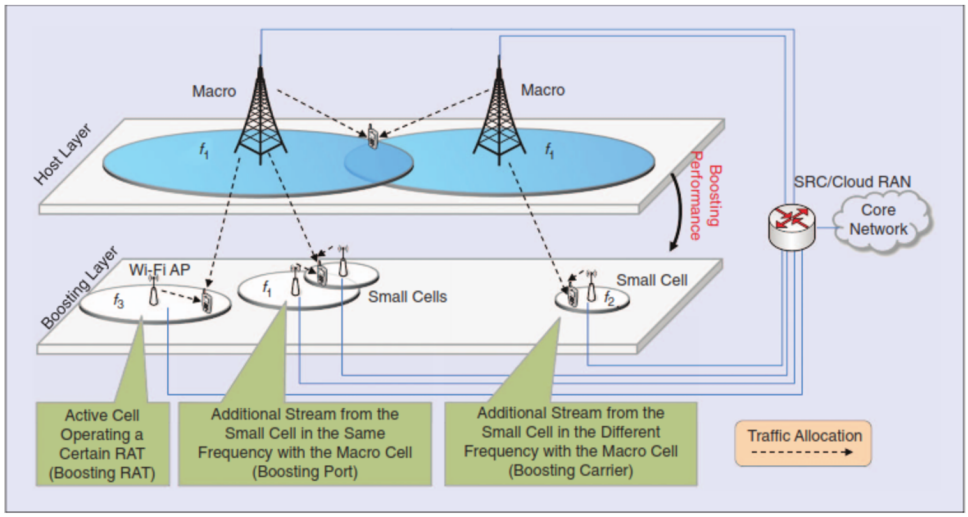


Figure 9.14: The role of essential intelligence for heterogeneous networks [15]

9.2.3.5 Cloud RAN (C-RAN)

C-RAN takes over the concept of a centralised baseband pool, where a lot of remote radio heads (RRH) connect to the center of a baseband units (BBUs), see figure 9.15. With this a maximized spectrum efficiency can be achieved through the cooperative radio with distributed antennas equipped in RRHs.
NFV can be a solution to create the virtual base stations for the cloud infrastructure of C-RAN. And the remote radio units (RRUs) can be implemented by SDR based on an open platform. But if multiple standards are allocated to the spectrum, RRH can handle this only partially. When the standard is changed the BBU is forced to restart, instead of sharing multi-standards resources directly. In heterogeneous networks (see 9.2.3.4) Cloud-RAN can achieve low complexity and energy efficiency through shifting the data processing from Base Stations to virtual Base Stations [28].

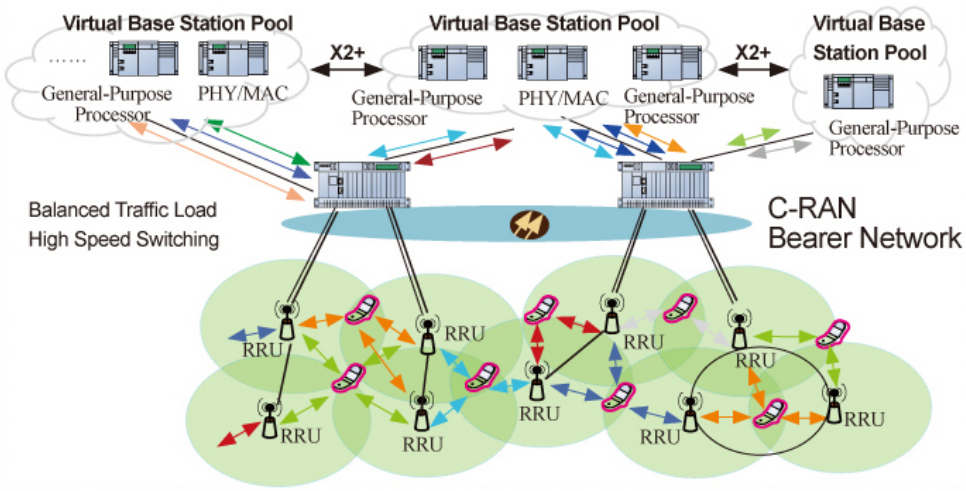


Figure 9.15: C-RAN [1]

9.2.3.6 Automated network organization: Unified self-management architecture

To manage future complex 5G networks automation appears to be the only cost-efficient way. Self-organized networks (SON) with self-configuration, self-optimisation and self-healing, SON self-optimisation is currently used in 4G to reduce manual involvement to improve mobility robustness, coverage and capacity optimisation, mobility load balancing and energy saving. These SON have to be improved to a unified self-management architecture for 5G systems [23].
The SEMAFOUR unified self-management system achieves coordinated and focused operation of SON through integrated SON management system, multi-RAT/multi-layer SON functions and decision support system. The integrated SON management system consists of policy transformation and supervision, operational SON coordination and monitoring functions. This turns network-oriented objectives into exact execution policies and rules for individual closed-loop SON functions. SEMAFOUR is one of the strong competitors for future SON evolution, while for future unified self-management systems the SDN concept should be considered [23].

9.2.3.7 Improved network architecture and deployment

Figure 9.16 shows a hybrid architecture of NFV, SDR and SDN. Which is focused on the functionalities required for the virtualization and the following tasks of an operator. NFV can increase the flexibility of network service deployment and integration in an operator's network. To achieve the NFV goals it makes use of the SDN mechanisms, enhancing performance, simplifying compatibility with existing deployments, and facilitating operation and maintenance procedures. In return, NFV provides the infrastructure upon which the SDN can run. And SDR provides function virtualization support for mobile networks [28].

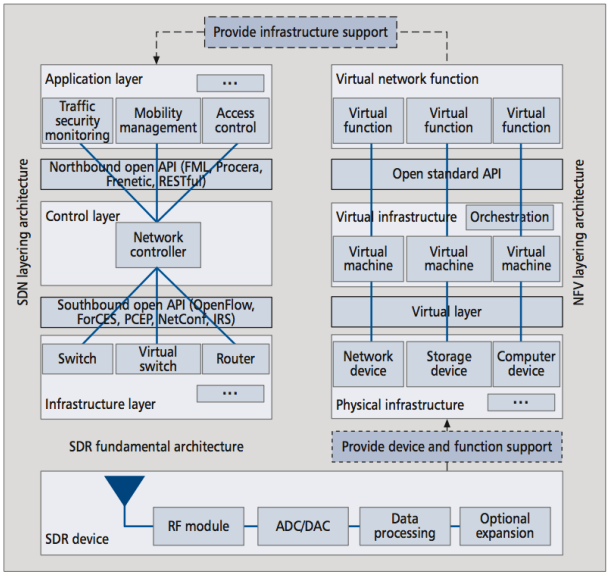


Figure 9.16: The architecture of NFV, SDR, SDN [28]

Hardware capacity can be expanded and complexity can be reduced through logical separation of software and hardware through and virtualization [23].

9.2.3.8 Smart devices

For the 5G connectivity in D2D (see 9.3.2) it is mandatory that smart devices are more intelligent, which also implies an architectural change. Additionally it should be capable of connecting to several wireless technologies simultaneously. Moving with vertical handovers requires radio access to different licensed or unlicensed bands. Further on the 5G smart devices are reconfigurable, multimode, cognitive radio-enabled. The reconfigurability can be achieved by SDR (see 9.3.2) concept [25].

9.3 Economical Perspective

In this section, effects of discussed technologies on cost and quality of service (QoS) are discussed. As costs can inhibit the implementation of any technology, costs have to be kept as low as possible for new technologies such that they are adopted by major cellular network providers. It is also important to differentiate between costs beared by cellular network providers and costs beared by users. Cellular network providers will have an incentive to push technologies for which users will have to pay directly. Main cost drivers are hardware costs, energy consumption of networking equipment as well as licensing fees. Besides cost, new business models arise. Figure 9.17 gives an idea about what opportunities of 5G could be.

Figure 9.17: Business opportunities of 5G [4]

9.3.1 Evolution of Radio Access Technologies

The main issue addressed by 5G from an economical perspective is the growing demand for fast mobile internet. Furthermore, the growing number of customers who are using the internet from mobile devices clearly shows that this use of the internet will be increasingly important in the future. From the literature review of the previous chapter, performance, robustness and reliability are important aspects from a customer's perspective, which should be taken in the development of 5G. Furthermore, it can be assumed that many customers wish to be able to use their mobile devices for a longer period of time without having to charge their devices batteries. From an economical perspective, these customer needs should be seen as key drivers in the development of 5G. 5G can be seen as an answer to the growing demand of bandwidth and mobile spectrum. mmWave enhances bandwidth capacity and Cognitive Radio helps taking advantage of the existing spectrum to a higher extent. 5G also delivers a solution for higher data rates:

- **Asymmetric traffic** in full duplex radio allows maximizing data traffic
- **NOMA/SCMA** allow better performance in multiplexing
- **Massive MIMO** increases the throughput
- **Ultra Low Latency** is in favor of user experience and allows reasonable performance in M2M communications

Higher data rates allow new business models, e.g. data intensive applications including video or games, which may increase the revenue of all major economic actors in the field. In terms of feasibility of 5G, the presented technologies are, partly, still in development and have to be standardized, which will reduce cost of deployment. However, many of the consulted studies show that applying the technologies may be possible within a reasonable amount of time. In Massive MIMO it can be seen as an advantage that the number of base station antennas is scalable. However, due to low beams in the context of mmWave antennas, there are still concerns, which can burden the application of this technology. In times where questions about the future of many countries in terms of energy supply are on the rise, it is also worth considering energy efficiency of new technologies. As network components in Massive MIMO tend to low-power, Massive MIMO contributes to energy efficiency and, thus, savings in cost. The presented design principles of 5G (e.g. "always available" instead of "always on") also strive for more energy efficiency. In terms of robustness, reliability, which are two criteria determining the usefulness of a technology and are thus economically relevant, 5G also provides progress. Massive MIMO exhibits high robustness against intentional jamming. DSA allows predictable quality of service,

which is an advantage for data based applications which have to run reliably. For instance, phone calls over the mobile data network could be expected to become more reliable with 5G.

9.3.2 Blended Infrastructures & Traffic Offloading

Effects of heterogeneous network deployments (see section 9.2.2.1) are manyfold. Chen et al. [11] mention a lot of positive and negative factors. First of all, due to the smaller cell size, the transmission power of small cells will be orders of magnitude less than that of today's macrocells. This provides more flexible spectrum usage and energy savings. Indoor small cells are currently built according to macro cell parameters, which in manufacturing costs a lot more. Optimization and adaption of the cells to their surroundings does not only yield higher performance, but also reduce the cost of integrated circuits [11] [15]. Currently, power efficiency is low in macro cells due to transmission loss. With small cells, transmission power can be reduced, and cooling systems are not necessary anymore. Additionally, inactive small cells can be deactivated, resulting in higher energy efficiency of the system. With the separation of user and control plane, traffic is automatically routed to small cells if available. This loads local small cells more, which themselves have the advantage of smaller transmission power and therefore lower energy consumption [11]. However, in most approaches, dual receivers and transmitters are needed inside an attached device. Overall, an increase in QoS can be expected, since the small cells deliver a greater coverage - even in indoor areas - as well as higher available bandwidth due to the usage of higher frequency ranges, able to transfer data more quickly. If femtocells are used instead of Wi-Fi hotspots, femtocells are also able to deliver speech transmission. Cost can be reduced by setting up femtocells in the first place [10].
Pro's and con's of composite wireless infrastructures (see section 9.2.2.2) are multifaceted. Current Wi-Fi implementations by cellular operators contain options to offload all data traffic to Wi-Fi. This implementation is cheap, and utilizes unlicensed spectrum, but the used MAC protocol is not well suited for heavy traffic. And White-Fi offers other, currently unlicensed frequency ranges that could be used [10]. For cellular network providers, Wi-Fi can therefore be a cheap option to provide cellular access for little to medium congested areas. Depending on the amount of data needed, UDRANET's could fit the role for high speed data transfers.
When talking about hybrid topology networking (see section 9.2.2.3), one of the biggest problems is that current mobile systems topology is inefficient, as all data has to go through the core network. Hybrid topology networking with technologies such as D2D communication and relay will allow offloading of data and make the whole system more efficient. In addition, the number of cells does not have to be increased to get a higher coverage, which saves on costs [11] . With D2D and relaying communications, if frequency spectra are shared between the cell and the devices, a lot of interference will be the consequence, while no new transmitter inside the devices is needed. Using a dedicated frequency allows the usage of high spectrum frequencies for higher transfer rates, needs a new set of receivers / transmitters though. And interference with the current network could be reduced drastically. In thought of the changeover from 4G to 5G, Chen et al. [11] predict that both frequency sharing and frequency split will be implemented to make the technology available for everyone in the first place, and later on make the change to frequency split.
When taking into consideration caching (see section 9.2.2.4), traffic offloading of the core network could be accomplished by proactive caching [7]. This would allow the usage of off peak hours to load presumably used data, based on training machine learning models. Its advantages would be a more flatted out consumption curve, where peaks in highly frequenced hours could be prevented. If rolled out, this would lower the cost, as possibly

fewer antennas would be used. In addition, QoS could improve, given that the machine learning models work, as also in by cellular networks uncovered areas content could be consumed.
All in all, economical benefits as well as QoS are going to increase with mentioned upcoming 5G technologies. By suiting cells on their application area, the cost of a newly deployed cell can be reduced. Using D2D communications will reduce cost for cellular operators as well, as existing devices could be used as relay antennas. On the other hand, relay and D2D technologies will require additional receivers and transmitters in user devices, which will generate additional cost on the user's side. Depending on how the frequency spectrum is licensed in a country, costs might increase due to broader frequency spectra needed to support different frequency spectra for different cell sizes. These new frequency ranges also require new transmitters and receivers on the side of cellular network providers as well as the side of users. By adjusting transmission power depending on application of a cell, no longer needed cooling systems due to smaller transmission power and deactivation of temporarily unused base stations, the energy consumption can be reduced.

9.3.3 Introduction of Intelligence

From an economical point of view, SDR contributes to cost efficiency in several different ways. As software is expected to take over some hardware functionality, expensive hardware (like filter, modulation demodulator, and converter) no longer has to be acquired. This also allows operators to quickly adapt to changes in the radio frequencies, without distributing new hardware. Often, different components of hardware are from different companies and might cause compatibility issues, which even further increases cost. This is why technologies have to be standardized. Furthermore, cost and space savings on the user devices are expected, because more general purpose hardware and less specific hardware is applied. Updates in usage rules for the different frequencies are possible through chips on the user devices. However, one hardware upgrade to the latest standard modules might be required to receive and interpret radio [12].
Cho et al [12] claim that the waste of network utilization should be reduced. Otherwise service providers may be struggling with hardware costs that exceed the benefits of the high bandwidth requirement of 5G. This appears to be especially relevant considering that Bastug et al [7] expect that, in 2020, the volume of mobile will have increased by 1000 times (compared to today) and 10-100 times more devices will be connected.
Through SDN, the bandwidth utilization rate can be greatly increased. This may result in a huge cost benefit. For example, Google saved 2/3 of the hardware cost for the same QoS as they had before implementing SDN [12]. Due to abstraction, SDN supports creating, testing and deploying new network applications from weeks or months to hours or days [15, 17]. As stated in the ONF White paper [17], SDN technologies address challenges of high bandwidth, the dynamic nature of applications, network adaption to business needs and significantly reduce the operations and management complexity.
Benefits of NFV include the reduced power consumption and Hardware costs [15]. Additionally, services can be created, tested and deployed in a shorter amount of time (compared to before) and are easier to manage. Network infrastructure is also easier to manage with NFV.
Fusion Net allows a network operators to apply the boosting layer where the traffic is generated, while the coverage is still supplied by the host layer. Fusion Net aims to avoid over-provisioning of resources and maximizes the energy and cost efficiency [15].
C-RAN enables the usage of processing power from cloud computing in order to process the data from RRHs. This contributes to wasting less processing power. Thus, hardware and cooling can be saved on the Base Station side and aggregated in Data Centers, which

can be used by multiple thousands RRHs. Therefore, C-RAN can lower the energy costs and the complexity in heterogeneous networks [30].
An approach for cost effective management of complex 5G networks is automation. A strongly automated network organization (including self configuration, self optimization and self healing) can greatly reduce manual involvement and, thus, reduce costs [23].

9.4 Social Perspective

In this section, impacts on quality of experience (QoE) of mentioned techniques are discussed out of a user's perspective. A new technology in 5G that has mostly negative effects on the user's experience should not be implemented. As ecological impacts are also of a user's interest, their impacts are also discussed here. Assumptions are referenced if possible, else are implications of the authors.

9.4.1 Evolution of Radio Access Technologies

From a user experience's perspective, 5G's improvements in performance and quality of service are clearly beneficial. The evolution of radio access technology does not force users to change the way they use their devices. However, improved bandwidth capacity and improvements in performance will enable new business models and add value to the way users use their mobile devices. Concretely, more data intensive applications like videos or games will benefit from 5G.
As described in the previous chapter, performance will be enhanced by following technologies: Asymmetric traffic with full duplex radio; mmWave, Massive MIMO. Users will appreciate this plus in mobile network performance. QoS will also be enhanced due to Massive MIMO (higher robustness using a high number of antennas) and NOMA/SCMA (providing more robustness in multiplexing).
Social concerns towards 5G may be connected with mmWace and Cognitive Radio. Due to low beams of mmWave antennas, environmental concerns may arise. Furthermore, Cognitive Radio enables secondary users to access spectrum over primaray users' devices. Some users may be cautious about this, because they fear that this kind of access is a disadvantage for their data security.
Due to predictable quality of service with DSA, users will be able to make phone calls over the mobile data network in a reliable way. This makes phone calls cheaper for users (as necessary resources can be utilized in a dynamic manner) and provides the same experience. Customers may also be glad about 5G's improvements in energy efficiency through new technologies and design principles which are in favor of saving energy, also in end user devices.

9.4.2 Blended Infrastructures & Traffic Offloading

When using heterogeneous network deployments (see section 9.2.2.1), due to the separation of user and control plane, hand-over between cells is quicker, as multiple cells are controlled by the same "bigcell". For the user, the switch between antennas will therefore be less noticeable in cases where switches inside bigcells happen [11], and QoE will therefore increase. All in all, through the different sizes of cells, coverage will get noticeably bigger for the user, and also the bandwidth will get bigger.
If being compared to composite wireless infrastructures, the advantage of having speech service indoors instead of only data access is also noticeable, which is currently not possible with composite wireless infrastructures.

For composite wireless infrastructures (see section 9.2.2.2), when integrating VoIP into cellphones directly, the user won't be able to differ between composite networks and heterogeneous networks, as with both data as well as speech can be handled.
Hybrid topology networking (see section 9.2.2.3) will bring better network coverage and higher possible bandwidth to every user, with the cost that every device in the network will act as node, and will therefore also consume battery if the user is not using his device. Battery life will therefore degrade more quickly, and battery technology has to be improved [25].
Caching (see section 9.2.2.4) will have influence on QoE, due to higher overall availability of the existing network thanks to fewer overall requests, as well as minimization of latency to an absolute minimum, as the data is already available on the users device. This has the negative effect, that storage space has to be allocated on the users device though, leaving him with less storage space. In addition, the machine learning models will need to gather quite a lot of personal user information to be able to predict accurately, and this could not be in the users interest.
All in all, the QoE will increase due to blended infrastructures and traffic offloading, mostly based on quicker handover between antennas, overall higher transfer rates and higher network coverage. Battery consumption of newer devices on the other hand will be degraded due to relay technologies.

9.4.3 Introduction of Intelligence

In general, the introduction of intelligence results in an improvement of QoE, being beneficial for the user. This improvement can be achieved by lower latency, more energy efficiency and substitution of hardware by software. For the most part, this leads to a reduction in costs for the same or even a better performance. In what follows, benefits for the specific technologies are explained.
As there is some research on possible connections between radiation and diseases like headaches, insomnia or cancer, SDR may mitigate the danger potential. SDN infrastructure can better adapt to dynamic user needs by checking bandwidth availability and adjust the data to the available bandwidth [17]. For example, the resolution of a video could be determined by the application. This could minimize delays and interruptions (which would decrease QoE). NFV also has a potential of reducing the cost for users and can open an opportunity to use hardware resources, as IaaS, through virtual networks.
Due to the performance boost, Fusion Net can increase QoE by a great extent. Users receive the best performance possible through the boosting layers. With C-RAN the user can experience lower latencies due the processing power and scalability of the Cloud. The user may also get a desirable cost reduction, because the network operator can save expenses for hardware and energy. The automated network organization does not impair the user's experience, except for maybe getting the same service for smaller costs.

9.5 Conclusion

The goal of this paper is to analyze possible upcoming technologies for 5G, based on state of the art research, and giving an outlook on three different perspectives. The perspectives are the technological perspective, where the different technologies have been explained and described, the economical perspective, where QoS and cost and benefit were outlined, and the social perspective, where QoE and effects on the user were reviewed. In the previous sections of this paper, we therefore gave an overview over the technologies and their effects on upcoming 5G. When solely talking about technology, researchers have

already outlined the upcoming 5G network quite well. However, the technologies still need to be standardized. In what follows, we summarize the most important developments.

- On one hand, it is clear that there is need for a wider frequency spectrum. Technologies supporting this claim are mmWave (see section 9.2.1.2), Cognitive Radio (see section 9.2.1.5), SDR (see section 9.2.3.1) and Cell shrinking in form of Heterogeneous Network Deployments (see section 9.2.2.1).

- On the other hand and concluding, radiation levels will overall increase due to these technologies, which might have effects on the environment. However, there also exist technologies reducing radiation levels, such as offloading traffic from macrocells to femtocells, where transmission power of macrocells can be reduced.

- Overall performance will increase due to multiple factors. First of all due to higher coverage due to D2D communication (see section 9.2.2.3), cell shrinking (see section 9.2.2.1), cognitive radio (see section 9.2.1.5) and Fusion Net with additional boosting layers (see section 9.2.3.4). Secondly due to technologies providing higher throughput, based on new transmission techniques as Asymmetric traffic & full Duplex Radio (see section 9.2.1.1), better usage of the bandwidth SDN (see 9.2.3.2) or the usage of higher frequency spectra with mmWave (see section 9.2.1.2) Thirdly lower latency by reducing the number of hubs the data flows through (see section 9.2.1.4) and finally a larger frequency spectrum and therefore less interference (see section 9.2.3.1).

- Higher energy efficiency of hardware due to design principles (see section 9.2.1.8), NFV (see section 9.2.3.3), C-RAN (see section 9.2.3.5) or substitution of hardware with software in SDR (see section 9.2.3.1), SDN (see section 9.2.3.2) and lower transmission power of antennas due to Heterogeneous Network Deployments (see section 9.2.2.1).

- Traffic will be offloaded of the main cellular backbone by multiple technologies, such as relaying and D2D and M2M communications (see section 9.2.2.3) and caching (see section 9.2.2.4).

Through overall increased performance of the 5G network, new business models and markets will be opened. When comparing the switch between 3G and 4G to the upcoming switch to 5G, there most possibly will not be a standard replacing LTE, which saves on cost. In the following, cost drivers, saving potentials and the QoS are discussed:

- Cost drivers include costs for new smallcells, multiple receivers and transmitters per mobile device, and cloud services.

- Saving potentials include the more efficient usage of the existing spectrum through Cognitive Radio (see section 9.2.1.5) and Dynamic Spectrum Access (see section 9.2.1.6), VoIP communications with technologies such as White-Fi (see section 9.2.2.2), better bandwidth usage with SDN (see section 9.2.3.2), Hardware cost reduction in SDR (see section 9.2.3.1), NFV (see section 9.2.3.3) and energy savings through higher energy efficiency (see section above).

- QoS will increase due to higher coverage, higher throughput and lower latency (see section above).

From a social perspective, the following effects from a switch to 5G can be expected:

- Thanks to the higher QoS, the performance of 5G will be noticably higher.

- New applications are possible due to the higher performance of the new 5G network, in areas such as gaming and video transmissions.
- The quality of voice transmissions will improve, thanks to technologies such as VoIP (see section 9.2.2.2).
- Battery life degradation through Relay technologies (see section 9.2.2.3), but on the other hand improvements due to a general increase in battery life of smartphones, as well as more energy efficient design principles.
- A quicker handover between antennas is possible thanks to technologies as Spectrum Mobility (see section 9.2.1.5) and separation of user and control plane (see section 9.2.2.1).
- As mentioned above, higher network coverage will be possible.

Concluding: 5G will tackle some major challenges we face today. The higher performance of the network will be able to withstand the upcomming rise in mobile and IoT devices for a reasonable amount of time. In addition, todays network congestion could be reduced thanks to traffic offloading strategies, as well as dynamic frequency spectrum allocation, tackling todays shortage in frequency spectrum.

Bibliography

[1] 2011.

[2] Wireless innovation forum's comments, international telecommunication union, 11 2013.

[3] 2016.

[4] Online Magazine 5G.co.uk. What is 5g? *5G.co.uk*, 2016.

[5] Ali Mohseni Ahouei. Massive mimo!!! how?? *LinkedIn*, 2015.

[6] Won-Yeol Akyildiz, Ian F. an Lee, Mehmet C. Vuran, and Shantidev Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks*, 50(13), 2006.

[7] E. Bastug, M. Bennis, and M. Debbah. Living on the edge: The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, 52(8):82–89, 2014.

[8] E. Bastug, J.-L. Guénégo, and M. Debbah. Proactive small cell networks. *20th International Conference on Telecommunications (ICT), Casablanca, Morocco*, 2013.

[9] R. Chávez-Santiago, Yoram Haddad, Lyandres Vladimir, and Senior Balasinghamm, Ilangko. Voip transmission in wi-fi networks with partially-overlapped channcels. *IEEE Wireless Communications and Networking Conference*, WCNC 2015, 2015.

[10] R. Chávez-Santiago, M. Szydelko, A. Kliks, F. Foukalas, Y. Haddad, K.E. Nolan, M.T. M.Y., Masonta, and I. Balasinghamm. 5g: The convergence of wireless communications. *Wireless Personal Communications*, 83(3):1617–1642, 2015.

[11] S. Chen and J. Zhao. The requirements, challenges and technologies for 5g of terrestrial mobile telecommunication. *IEEE Communications Magazine*, 52(5):36–43, 2014.

[12] H. H. Cho, C. F. Lai, T. K. Shih, and H. C. Chao. Integration of sdr and sdn for 5g. *IEEE Access*, 2:1196–1204, 2014.

[13] L. Csurgai-Horváth and J. Bitó. Primary and secondary user activity models for cognitive wireless network. In *Telecommunications (ConTEL), Proceedings of the 2011 11th International Conference on*, pages 301–306, June 2011.

[14] Mickael Dardaillon, Kevin Marquet, Tanguy Risset, and Antoine Scherrer. Software defined radio architecture survey for cognitive testbeds. *Universite de Lyon, Inria*, 2013.

[15] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xion, and J. Yao. 5g on the horizon: key challenges for the radio-access network. *IEEE Vehicular Technology Magazine*, 8(3):47–53, 2013.

[16] Ericsson. 5g energy performance. *Ericsson White Paper*, 2015.

[17] Open Networking Foundation. Software-de ned networking: The new norm for networks. 2012.

[18] Yoram Haddad and Dana Porrat. Femtocell: Opportunities and challenges of the home cellular base station for the 3g. Algarve, Portugal, 6 2009.

[19] Erik Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas Marzetta. Massive mimo for next generation wireless systems. *IEEE Communications Magazine*, 52(2), 2014.

[20] William Lehr, Amparo Canaveras, Raffaele de Peppe, Peter Ecclesine, Hannu Flinck, Bill Lehr, Prakash Moorut, Karen Sollins, Max Solondz, Martin Weiss, and Seppo Yrjola. Toward more efficient spectrum management. Master's thesis, MIT, Spectrum Policy Working Group, 2014.

[21] V. Letter, M. Kafsi, and E. Kazemi. "been there, done that: What your mobility traces reveal about your behavior," in mobile data challenge by nokia workshop. *Int. Conf. on Pervasive Computing*, 2012.

[22] Kai Lin, Wenjian Wang, Xianbin Wang, Wen Ji, and Jiafu Wan. Qoe-driven specrum assignment for 5g wireless networks using sdr. *IEEE Wireless Communications*, 2015.

[23] Z. Ma, Z. Zhang, Z. Ding, P. Fan, and H. Li. Key techniques for 5g wireless communications: network architecture, physical layer, and mac layer perspectives. *Science China Information Sciences*, 58(4):1–20, 2015.

[24] H. Malik, M. Ghoraishi, and R. Tafazolli. Cross-layer approach for asymmetric traffic accommodation in full-duplex wireless network. In *Networks and Communications (EuCNC), 2015 European Conference on*, pages 265–269, June 2015.

[25] L. Pierucci. The quality of experience perspective toward 5g technology. *IEEE Communications Magazine*, 22(4):10–16, 2015.

[26] Hien Quoc Ngo. Massive mimo: Fundamentals and system designs. Master's thesis, Linkoping Studies in Sicence and Technology, 2015.

[27] M. Rebato, M. Mezzavilla, S. Rangan, and M. Zorzi. Resource sharing in 5g mmwave cellular networks. *arXiv preprint arXiv*, 1603.02651, 2016.

[28] Songlin Sun, Michel Kadoch, Liang Gong, and Bo Rong. Integrating network function virtualization with sdr and sdn for 4g/5g networks. *IEEE Network*, 2015.

[29] Abd-Elhamind M. Taha. Green wireless networks: A radio resource management perspective. *IEEE ICC 2012 Workshop on Cognitive Radio and Cooperation for Green Networking*, 2012.

[30] Telefonica and Ericsson. Cloud ran architecture for 5g. *A Telefónica White Paper Prepared in collaboration with Ericsson*, 2015.

[31] Philipp Tracy. What is mm wave and how does it fit into 5g? *RCR Wireless News*, 2016.

[32] Gonzalo Vazquez-Vilar. Interference management in cognitive radio networks. *University Carlos III de Madrid*, 2016.

[33] Marsha Walton. Is wifi on steroids really the next big thing?, 3 2006.

Chapter 10

Comparison of Business Model Frameworks for the Internet of Things

Matthias Diez, Christian Ott, Silas Weber

With the emergence and spread of the 'Internet of Things' (IoT), businesses face a new challenge to stay productive and exploit the new technologies in a changing technological environment. Due to the inherent networked nature of IoT, businesses need to apply adapted business models in order to prevail in the IoT market.
In this paper, the way how business model frameworks have changed due to the ongoing developments in the IoT industry compared to traditional industries is analyzed. This is done by first giving an introduction into IoT. Then, traditional as well as specific IoT business models frameworks are presented. The core of this seminar report is the comparison of the discovered IoT business model frameworks by applying them to a fictional use case. It results that the reviewed IoT business model frameworks tend to shift their focus towards value creation and collaboration. Further, it shows that the ecosystem framework is yet in its infancy and is difficult to apply to the fictitious use case. However, the three-dimensional collaborator framework, combining the ecosystem and traditional canvas view, worked well for this use case and proves to be a ready-to-use framework for IoT businesses today.

Contents

10.1 Introduction

Introduction of technology innovation always leads to innovation in business models. With the emergence of the 'Internet of Things' (IoT), business models become possible that were unthinkable in the past. Developing business models in an emergent technology field can be challenging. Business model frameworks can help stakeholders to come up with better economic solutions for their business idea. Therefore it is important that frameworks adapt to technical and economical changes in their respective business environment.
The goal of this work is fourfold: First, a common ground should be built, with definitions and characterizations of IoT, business models, and business model frameworks. Second, the basic business model framework concepts 'Magic Triangle' and 'Business Model Canvas' should be introduced. Third, it should be examined what changes were made to business model frameworks due to the development of IoT, and lastly, the currently available IoT business model frameworks should be presented and compared against each other.
This seminar report is therefore structured as follows: In Section 10.2, the term 'Internet of Things' is introduced and its definition, growth, architecture, and potential of IoT for business is evaluated. Section 10.3 presents the concept of business models and business model frameworks. Along with the definitions, two major 'classic' business model frameworks are described. To emphasize the difference between traditional and IoT-specific business models, two predominant IoT business model patterns are presented in Section 10.4. Section 10.5 then goes on to show available business model frameworks that are adapted for IoT. After that, those frameworks are compared using a fictitious company as an illustrative example in Section 10.6. In Section 10.7, the findings of the previous sections are evaluated. Finally, a summary of the report as well as a conclusion about the current state of IoT business model frameworks are presented in Section 10.8.

10.2 Internet of Things

The term 'Internet of Things' has grown to be a major topic in academia and industry [18]. The phrasing was first used by Kevin Ashton during a presentation in 1998 [36]. From there, IoT has become a new paradigm postulating the "interconnection of physical objects, by equipping them with sensors, actuators and a means to connect to the Internet" [8].
Even though the term 'Internet of Things' is currently used by everyone, there is no common consent about its definition. The International Telecommunication Union (ITU) defined IoT as "a global infrastructure for the information Society, enabling advanced services by interconnecting (physical and virtual) things based on, existing and evolving, inter-operable information and communication technologies" [17] in a recommendation published in 2012. Besides ITU's view of IoT, other definitions were proposed. Atzori et al. [3] identified three visions of how IoT may be seen. As illustrated in Figure 10.1, one vision focuses on the 'things' being connected over technologies like radio-frequency identification (RFID), Wireless Local Area Network (WLAN) or Near Field Communication (NFC). The second, 'Internet-oriented' view envisions anything being connected with anything. The third point of view is a 'Semantic-oriented' vision. Definitions found in literature predominantly focus on this vision. The 'Semantic-oriented' vision present thoughts on problems generated through the extremely increased number of connected things, such as the challenging issues related to the representation, storage, interconnection, search and organization of information from IoT. Due to the differences in these visions, the definitions of IoT established by organizations or entities (see [3] Section 2) vary strongly depending on their specific interests, approach taken on the subject as well

as their backgrounds. The convergence of these three visions illustrated in Figure 10.1 can be seen as the overall paradigm of IoT. This converged perspective is the one that this report takes on in the remaining sections.

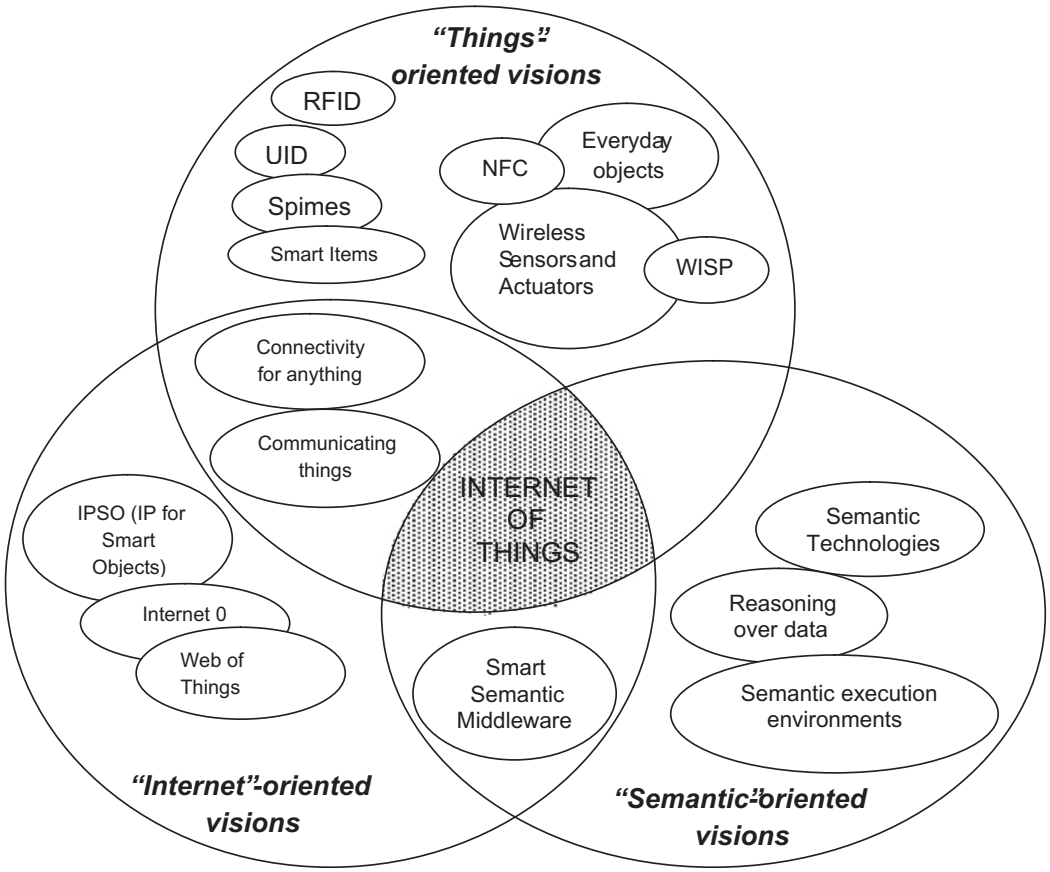


Figure 10.1: IoT paradigm as a result of the convergence of different visions [3]

The number of worldwide connected 'things' is rapidly growing. In 2016, Gartner predicted up to 6.4 billion connected devices which will rise up to 20.8 billion by 2020 [11]. Consequently, organizations are more and more involved with IoT-related products, applications and services either in the development or as an investment [18]. Google spent USD 3.2 billion to acquire Nest, a quickly growing smart thermostat company [34]. Samsung bought SmartTings, an open smart home platform [35]. Many IoT companies such as LIFX, a smart bulb company from San Francisco, are emerging at the moment. LIFX was funded thanks to KickStarter [19], a crowd-sourcing platform, and is today fully integrated with other smart home providers like the previously mentioned Nest and SmartThings. LIFX also works with Amazon Echo [2] or IFTTT [16]. Both merge a diversity of services and provide a single point of access. Telecommunication organizations invest in future technologies such as 5G which was discussed in Section 10 of this seminar report. Governments increasingly acknowledge the importance of IoT. The United States supported a Smart Cities Initiative with USD 160 million [24], and the Korean government planned investments of USD 5 billion in various IoT projects until 2020 [7]. The investments driven by national institutions as well as aggressive investments by companies in IoT are expected to "create new business opportunities and substantial social and economic benefits" [18].

Even though there is no concrete definition of IoT, there is some consent about the underlying architecture of IoT. A three-layered application stack can be found in various research literature. Albeit the layers may be named differently or be subdivided into more layers, they are semantically comparable [10] [18], [37]. In Figure 10.2, the layered approach from Wortmann et al. is illustrated [37]. They presented a classical three-layered

approach based on Porter and Heppelmann [29]. Only the 'application layer' is called 'IoT Cloud' instead.

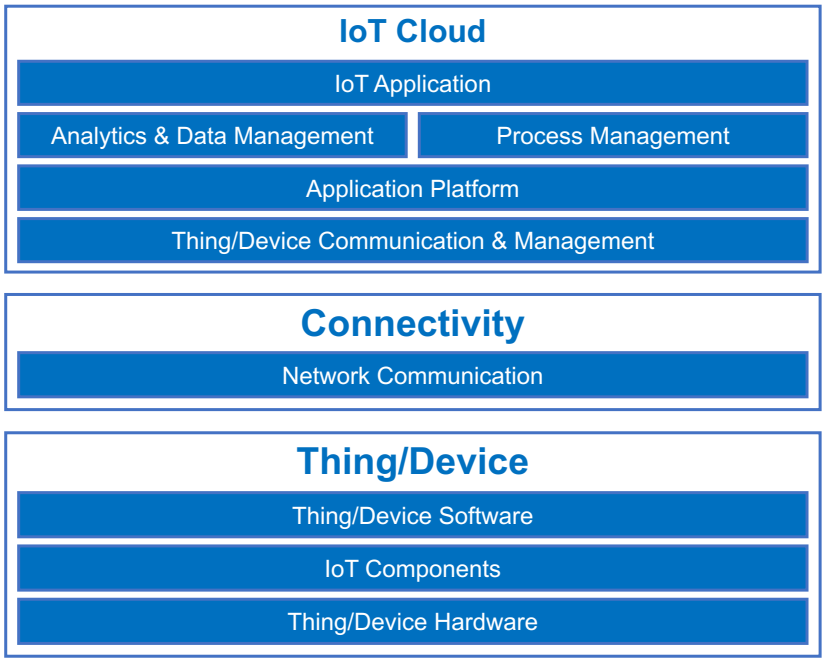


Figure 10.2: 'Internet of Things' technology stack by Wortmann et al. [37] (own illustration)

The 'thing/device' layer serves as basis. This hardware-based sensing layer has the function of identifying objects and collecting information through sensors over short-range and local networks [18]. The 'network' (or 'connectivity') layer transmits real-time data. It does not only connect people to 'things' but also allows an information flow between 'things' autonomously. The data presented by the 'network layer' can be used by companies to provide optimized and personalized services, which make up the top layer [18]. This 'application' (or 'IoT Cloud') layer is described by Ju et al. as a "combination of data processing and intelligence analysis to meet the industry needs to realize an intellectualized industry" [18] and allows companies to achieve "different types of intelligent application solutions" as well as the determination of business strategies [18].
The IoT has a vast potential. Customers enjoying novel experiences through the newly connected world on the one side and businesses getting involved in the IoT market with hopes of economical and reputational gains on the other side exist [18]. These 'gains' on the business side, respectively, the way to generate economical 'gain' in the IoT and how they have changed compared to value creation in traditional businesses, are the focus of the following sections.

10.2.1 Current State of the IoT Ecosystem

In this section, some of the relevant findings from the Cisco IoT ecosystem study are highlighted [6]. The following three topics are of special interest, namely players, market, and standardization in regard to the current state of IoT.
The following four categories of players have been identified by Cisco [6]:

1. **Commercial players in the offline world** are mainly manufacturers producing smart appliances such as smartphones, wearables or home automation devices, but also smart automobiles etc.

2. **Commercial players in the online world** are companies providing IoT-enabling services. Example companies are Amazon (Amazon Web Services), Google (Google

Cloud Platform) and Microsoft (Microsoft Azure) which all offer platforms and services that target the IoT market.

3. **Research and academia** will further incite the growth of IoT by creating new theories, products and materials. New innovations emerge from these players that disrupt the market.

4. **Governments and utilities** are the final group of players. They are the driving force for creating smart cities, smart grids and other smart infrastructure initiatives. They are the decision makers when it comes to regulation and policies for IoT technology and also play a key role in funding IoT developments.

There are numerous markets for IoT. Besides a market for consumer goods, such as smart phones, smart homes and other smart appliances, there are many more markets where IoT has a promising future [6]:

- **eHealth** includes such things as virtual health care, bioelectronics sensors such as heart monitoring implant, real time monitoring of vital signs and fitness monitoring.

- **Transportation** ranges from smart public and private transport to other smart transportation systems such as services like Uber, Lyft, and other car-pooling appliances.

- **Energy distribution** is expected to become more efficient and more resilient on supply and demand side through automation of the energy grid (smart grid).

- **Smart cities** are based on the idea of creating network infrastructures by using information and communications technology (ICT), to improve efficiency as well as development of all areas of urban life, including social, business and cultural services.

- **Manufacturing and distribution/logistics** are being transformed due to the current trend of IoT. 'Industry 4.0' or the 'fourth industrial revolution' refer to the automation and data exchange based on cyber-physical systems and cloud computing, creating a 'smart factory'.

- **Public safety** includes early-warning systems in smart cities. This helps preventing and mitigating catastrophes, and supports road-traffic safety measures and emergency medical services.

- **Agriculture** includes natural-resource management by GPS mapping technologies and sensors for analyzing crop yields, monitoring growth, measuring nutrient needs and similar.

- **Big data analytics** help to process the vast amount of data collected by all those sensors and smart devices. Properly analyzed, this data can deliver great insights for their respective applications. This is why big data analytics will immensely benefit from IoT but contribute to its development as well.

On the topic of standardization, Cisco found that current IoT-specific standardization activities are confined to single scopes of application, such as health care or agriculture. Those so called 'verticals' "[...] represent islands of disjointed and often redundant development" [6]. Such a fragmentation is detrimental to the IoT market. Overcoming and preventing such fragmentations poses the challenge of standardization. There are some major standardization bodies active in IoT such as the Institute of Electrical and Electronics Engineers (IEEE) or the World Wide Web Consortium (W3C). Not all of the standardization bodies have a global impact. For businesses, application standards will

enable interoperability between products. This is needed for businesses that are part of an ecosystem and rely on interoperability. Furthermore, standardization improves flexibility and prevents IoT businesses from becoming locked-in in proprietary industry standards. This flexibility increases competition between companies, which will in the end be beneficial for the end user and society, as it leads to better products at lower prices.

10.3 Business Model Frameworks

A successful business needs to offer products or services that there is a demand for and thus can be sold to make a profit. In short, a business has to create value. A company needs to have a strategy, a way of conducting business on an operational level that will result in long term financial success, otherwise a business is not viable and will disappear from the market. A business model could be seen as a logical abstraction describing how the business operates to make a profit. A business model hence is a core factor in deciding whether a business will be driven out of market or begins to prosper.
A **Business Model** is defined by Osterwalder et al. as "a description of the value a company offers to one or several segments of customers and of the architecture of the firm and its network of partners for creating, marketing and delivering this value and relationship capital, to generate profitable and sustainable revenue streams" [26]. The concept of the business model became important in the 1990s as the Internet began to spread and went on to become a central backbone of today's global economy since then [38]. There is no commonly accepted view as to what the business models should include [25], [26], [31]. Achtenhagen et al. [1] stated that there has been a change from 'what business models are' to 'what business models are for'. In various literature, there seems to exist an agreement that a business model is 'the way of doing business' for a particular firm [36]. At the heart of each business model stands the goal of minimizing cost and maximizing revenue [18].

A **Business Model Framework** is a tool to help a company in the development of its business model by presenting an overview over the business model components described above [8]. Two well-known business model frameworks are described in this Section. In Subsection 10.3.1 the 'Magic Triangle' by Gassmann et al. [13] is presented, followed by Subsection 10.3.2 which describes the business model canvas by Osterwalder et al. [27].

10.3.1 Magic Triangle

Having a suitable business model is vital for its success. In fact, "[...] business model innovators have been found to be more profitable by an average of 6% compared to pure product or process innovators" [13]. With the spread of IoT, new business models need to be developed to adapt to those technological changes in the environment and seize the opportunities that arise with such a change. But there seems to be a problem. "Very few managers are able to explain their company's business model ad-hoc, and even fewer can define what a business model actually is in general." [13]. To keep it simple, yet sufficient, Gassmann et al.'s conceptualization consists of only four central dimensions: 'Who', 'What', 'How' and 'Value'.
The Who, What, How and Value make up the Gassmann et al.'s **Magic Triangle**. The 'Who' addresses the target customer group. The 'What' refers to the product or service to those customers.
The 'Value' explains how the business model is financially viable and how value is generated. It includes cost and revenue structures. The 'How' addresses the process and

activities as well as the resources and capabilities that are required for building and distributing the value propositions.

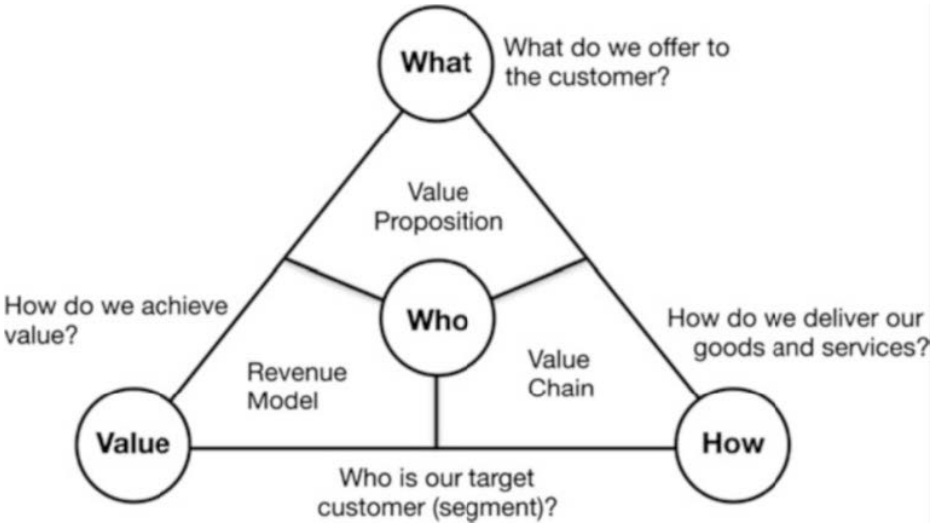


Figure 10.3: Gassmann's 'Magic Triangle' [13]

The constellation of the 'Magic Triangle' with the 'Who' node in the middle is shown in Figure 10.3. Gassmann et al. argued that reducing the business model to those four components allows for a simple yet thorough enough view of the business model architecture. But creating new business models is not an easy task. It requires thinking outside the box, going beyond conventional industry philosophy and can quickly become complex. Another problem is the 'not invented here syndrome', meaning that ideas, which are not from within the company, are disregarded solely because they come from the outside. As a consequence, business models should not just be copied from somewhere else but rather bring in external stimuli when generating ideas from within.
Gassmann analyzed 250 business models in different industries from the last 25 years and as a result identified 55 patterns of business models that have served as the basis for new business models in the past. Then, together with selected companies they developed a construction methodology based on the fact that 90% of all new business models have recombined previously existing ideas, concepts and technologies. The 'Business Model Innovation Navigator' is a ready-to-use methodology for coming up with new business models consisting of the following three steps:

1. Initiation: Describing the current business model is a good starting point. It creates a common ground for discussing the things that are done well and what needs improvement, and opportunities that are open to be exploited. Also, it gets the participants started thinking in the ways of business models.

2. Ideation: Recombining existing ideas helps generating new business models. For this purpose, Gassmann et al. condensed the 55 business models into a set of pattern cards. Each card has a title, a description and an example. The goal is applying different cards to the current model to see what would change in this situation. The cards should trigger discussions and act as a stimulus for new innovative ideas. Open-minded team members are essential, preferably from different functions. This allows different viewpoints and 'thinking outside the box' as well as overcoming the prevailing industry logic.

3. Integration: It is obvious, that a newly generated idea can not be implemented instantaneously. New ideas need to be gradually fleshed out into fully operational business models. Considering the new stakeholders, partners and consequences for the market is crucial.

10.3.2 Business Model Canvas

The Business Model Canvas (BMC) is another business model framework that helps to think about business models, independent of the industry the business is operating in. This strategic management tool was developed by Osterwalder in his doctoral thesis in 2004, and later published in book form by him, Yves Pigneur and Alan Smith [27]. The tool supports business developers in gaining an overview about the key factors of their business. The use of the BMC to develop and explain business models is widespread in literature and practice today. There are nine key factors explained by Osterwalder in detail, called 'building blocks'. All of the building block can be aligned next to each other in a way that visualizes the mechanics of the business model (i.e. the interactions between the building block) and makes it easier to understand and discuss it. The analog, pen-and-paper nature of the business model canvas allows it to be filled out in a team, stimulating creativity while finding a suitable business model for the business idea at hand.
The usual order of filling out the BMC, as shown in the video by Strategyzer [32], begins with the targeted **Customer Segments**, then following with the **Value Proposition** to these customers and the **Channels**, which allow the business to deliver the described value to the customer. The value is not only delivered in a physical way, but there is always an emotional **Customer Relationship** with the customer involved that needs to be considered. In exchange for the proposed value, customers are willing to pay money, which makes up the **Revenue Streams** of the business model. With this fifth building block, the money-generating revenue side of the BMC is complete.
As a counterpart to the revenue side, every business model also includes a cost side. In Osterwalder's canvas, the cost side consists of the following four building blocks: The **Key Resources** are critical to be able to deliver the proposed value. But resources alone do not create value - they need to be used for **Key Activities** in order for the business model to be viable. As a business owner, the decision of what part of the value you create yourself and what you let other companies do for you is written down in the **Key Partners** building block. The costs induced by resources, activities and partners are finally collected in the **Cost Structure** building block. All of these blocks are combined into one rectangular canvas as shown in Figure 10.4.

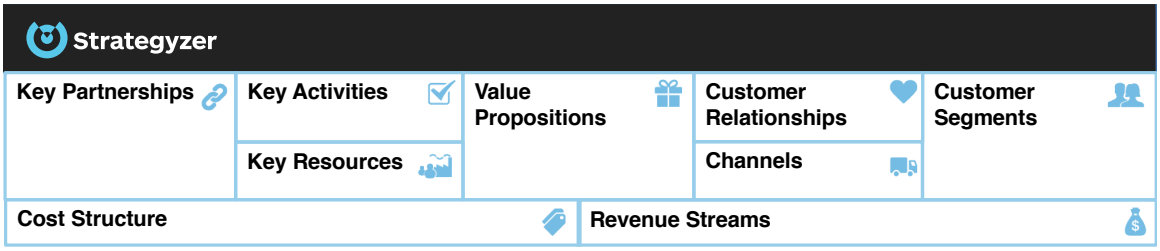


Figure 10.4: Business Model Canvas [32]

Each building block can be filled with different characteristics or types [27]. This gives credit to the idea that new business models are mostly recombinations of existing building blocks from other proven business models. The following paragraph will shortly summarize the options for each building block. A 'Customer Segment' can be a mass market, a niche market or a multi-sided platform. Business can have very diversified customer segments, with clear or unclear segmentation. It can be important to focus on the most important customers first and align the business model primarily for them. 'Value Propositions' can have various characteristics such as the newness, performance, cost reduction potential or accessibility of the proposed value. The key questions to answer with this block is 'What value is delivered to the customer?' and 'Which customer needs are being satisfied?'. There can be multiple different value propositions per company. A company usually offers multiple 'Channels' to serve their customers. Different customer segments

may require different channels to get the product or service, which the company offers. Also, the channel may differ along the different channel phases from Awareness, Evaluation, Purchase, Delivery to After sales. 'Customer Relationships' differ strongly from company to company. Some forms of business models rely heavily on a personal customer relationship (e.g., barbershop) whereas for other companies, the customers might not even be known because they serve themselves in a self-service manner. 'Revenue Streams' can stem from different sources and pricing models. A company can generate income through asset sale, fees for the usage, subscription, brokerage and licensing, or through advertising. Pricing models can be fixed per feature, customer segment or volume, or dynamic which means that every customer receives a different price depending on the current context. 'Key Activities' include all actions that are required to offer the 'Value Propositions', to run the different 'Channels' and maintain the 'Customer Relationships'. 'Key Resources' include the physical, intellectual, human and financial assets of the company. The company pools these resources in order to turn them into an added value, which can then be proposed to customers. The decision to include 'Key Partners' into a company's business model can have various reasons. It might be that the partner can provide a needed component of the business model which is cheaper, which includes less risk and uncertainty for the company or which can not be provided by the company itself. The 'Cost Structure' can be characterized along an axis from Cost-driven (lowest costs in the market) to Value-driven (best product/service in the market). Important characterizations of the cost structure include fixed and variable costs as well as the influence of economies of scale (the more you sell, the lower the cost per item) and economies of scope (the more diverse your offering, the more efficient a company can use its resources).

10.4 IoT Business Model Patterns

In this Section, the gap between the traditional business model frameworks described above and the available IoT business model frameworks, which will be presented in Section 10.5, is closed by having a closer look at two specific business model patterns (i.e. recurring types of business models) applied in the IoT environment as described by Fleisch et al. [10]. Their white paper is a starting point to find out what areas a relevant IoT business model framework has to cover. The analysis of 55 distinct business model patterns from Gassmann et al. [13] yielded two novel business model patterns, which can be applied to a IoT business idea: 'Digitally Charged Products' and 'Sensor as a Service'.
The pattern of **'Digitally Charged Products'** is described as "classic physical products are charged with a bundle of new sensor-based digital services and positioned with new value propositions" [10]. The components that can be used and combined for the creation of new business models are the following:

- Physical Freemium: Physical asset with free digital service, premium charged service offered.
- Digital Add-on: Cheap physical asset, digital services can be bought or activated at a high margin.
- Digital lock-in: Limit compatibility, high dependency, unable to use another service without high switching costs.
- Product as Point of Sales: Physical products become sites of digital sales and marketing services.
- Object Self-Service: 'Things' can place orders autonomously over the Internet.

- Remote Usage and Condition Monitoring: Smart things collect and send data in real time, which enables real time monitoring and error prevention.

This business model pattern with its six components embodies the "idea that the Internet of Things in its applications links digital services to physical products to create a hybrid bundle that is a single whole" [10].
The second pattern presented in the white paper is named **'Sensor as a Service'**. It is based around the business idea of "collecting, processing and selling for a fee the sensor data from one subsection to other subsections" [10]. The data gathered by the sensors itself is moved into the central focus, instead of the sensor or product itself as in the pattern 'Digitally Charged Products' described above. Instead of collecting the data for one application only, data can be sold in a multi-sided market to many different stakeholders. As pointed out above, Fleisch et al. do not deliver a business model framework but only two business model patterns ready to be applied to IoT business ideas.

10.5 Available IoT Business Model Frameworks

In this section, three different business model frameworks specifically designed for the IoT environment, which are currently available in research papers, are presented and discussed. Dijkman et al. suggest a framework based on the 'Business Model Canvas' discussed in Section 10.3.2. Westerlund et al. focuses more on the business ecosystems from a more abstracted point of view and Chan presents a three-dimensional collaborator model.

10.5.1 Adapted Business Model Canvas for IoT

Dijkman et al. [8] analyzed current research literature in order to create a business model framework for IoT applications. They searched for papers that contained the phrasing 'Internet of Things' together with 'business model'. Of the resulting 20 papers, only the five papers that contained actual business models were selected [4] [9] [21] [23] [33]. Two of the five papers were based on the 'Business Model Canvas' described in Subsection 10.3.2 [4] [33]. They identified the building blocks and building block types of business models through the selected papers and through interviews with professionals working in the IoT industry. Lastly, they determined the relative importance of each building block or type through a survey. The result of Dijkman et al.'s research can be seen in Figure 10.5. The building blocks observed were equal to the building blocks from the BMC, described in Section 10.3.2. In order to 'fill' the building blocks, the building block types presented by Osterwalder et al. were merged based on the interviews, divided into multiple types, removed or extended by additional types [28] [8]. The modified types are indicated in Figure 10.5, using a gray colored background.
Based on the interviews and a survey by Dijkman et al., the relative importance of each building block was determined. Dijkman et al.'s work showed that 'Value Propositions' is the most important building block in IoT business models. Furthermore, 'Customer Relationship' and 'Key Partners' were also considered to be more important by the interviewees. All other blocks had comparable importance results, with 'channels' being slightly less important. The framework presented by Dijkman et al. can be seen as an extended classical BMC [32]. Dijkman et al. took the BMC and filled it with building block types commonly used for IoT business models. Thereby, they helped the user of the framework by delivering main 'pillars' in the creation of a business model. These pillars can be seen as key parameters to lead the creation process into the right direction.

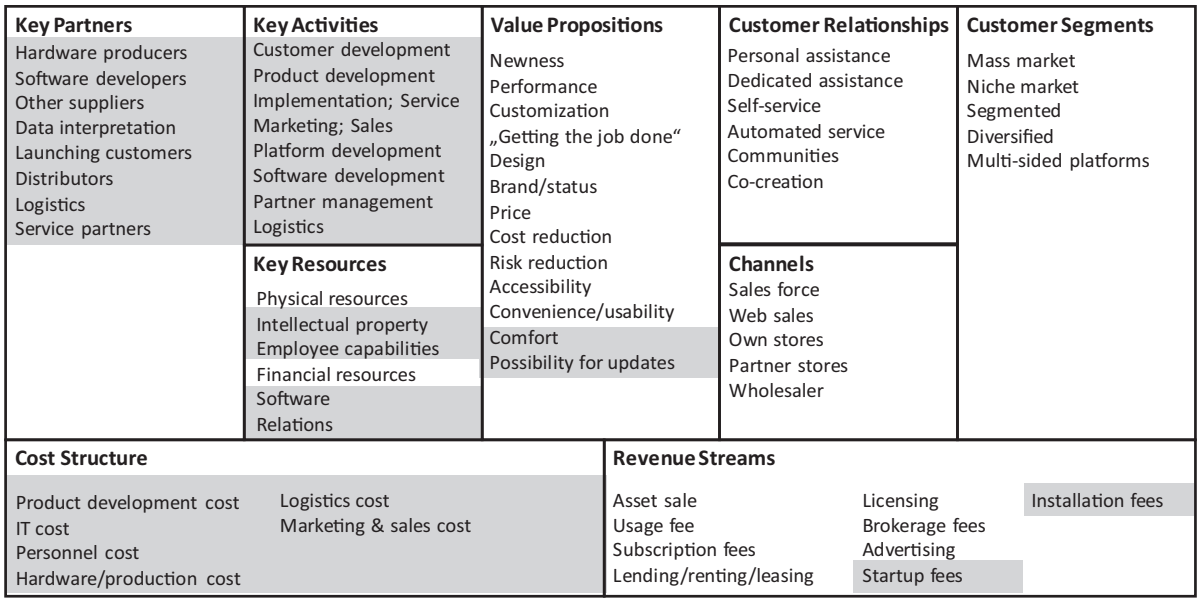


Figure 10.5: Business model framework for IoT applications by Dijkman et al. [8]

10.5.2 Ecosystem Business Model Framework

Westerlund et al. [36] move from seeing IoT mainly as a technology platform to viewing it as a business ecosystem. Thus, a shift from the traditional business model of a firm to designing ecosystem business models is postulated. Such an ecosystem business model is composed of value pillars that create and capture value. They identified three major challenges for designing ecosystem business models namely the diversity of objects, the immaturity of innovation in the field of IoT, and the unstructured nature of ecosystems in general.

As businesses move from centralized toward decentralized and distributed network structures, they become part of complex business ecosystems. A business ecosystem can be seen as an organization of economic actors. Those actors' business activities are anchored around a platform. It is argued that such systems are more than the sum of its parts and hence "operations of the system cannot be understood by studying its parts detached from the entity" [36].

Existing business model frameworks such as the 'Magic Triangle' and the BMC described earlier, are arguably not adequate enough when it comes to analyzing such ecosystems. With the emergence of IoT, the interdependency of actors in an ecosystem gains importance due to the networks inherent to such ecosystems.

As for all business, making money is essential. Three problems were identified when it comes to making money in IoT:

1. 'Diversity of objects' refers to the variety of different types of connected objects. Without a widespread standardization, it will be difficult to be efficient in an ecosystem. Managers will face a difficulty when trying to bring the objects, businesses and consumers together. Things can integrate with other things, requiring specific business logics.

2. 'Immaturity of innovation' refers to the current multitude of emerging technologies and components as well as innovations that have not yet matured into products and services. For IoT to be successful, modularized objects of 'plug and play' nature are needed.

3. With 'unstructured ecosystems', the problem of lacking governance and underlying structures is pointed out. Unstructured ecosystems may lack essential participants, for example an IoT operator or potential customers. New business opportunities arise when new relationships in new industries are built or when already existing connections are extended.

The ecosystem business model framework establishes a basis for building new business models to overcome these previously discussed problems and fit the ecosystem nature of IoT. The proposed framework consists of four key pillars, namely 'value drivers', 'value nodes', 'value exchanges' and 'value extract' as shown in Figure 10.6:

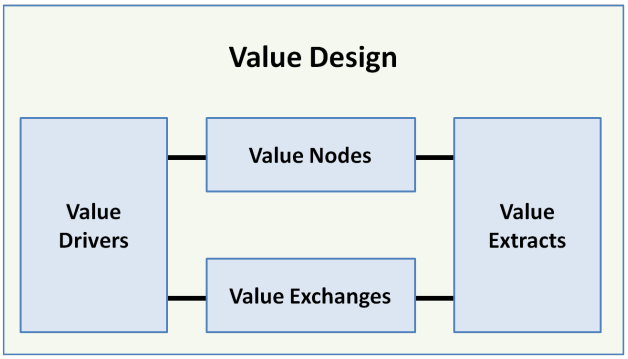


Figure 10.6: The four pillars by Westerlund et al. [36]

There are different **value drivers** in an ecosystem. Those value drivers are composed of both individual and shared motivations of the participants in the ecosystem. The shared motivations, i.e. the shared value drivers are crucial for creating a win-win situation in a trustworthy ecosystem. Without a long-term relationship between the actors built on mutual respect of their business goals, the ecosystem will fail. Each value driver also serves as an individual value node's motivational factor. Examples of shared value drivers are cyber-security and improved customer experience.
Value nodes consist of various actors, activities and processes linked with other nodes to create value. Further, these nodes also can have 'things' as actors. Here, smart things come into play. These self-acting 'things' may be sensors, smart machines, or other intelligence. Value nodes are heterogeneous: They can be organizations, networks of organizations or even networks of networks.
Value Exchanges take place between, but also within, different value nodes in the system. The value exchanged can be resources, knowledge and information. Thus, the value can be both tangible as well an intangible. Value exchanges are best described as a flow that powers 'the engine'. Value exchanges are important, because they describe how revenues are generated and distributed in the ecosystem.
As not all created value is meaningful in regard to commercialization, only specific parts of created value make sense to extract. **Value extract** refers to the part of the ecosystem that extracts value. That means it focuses on what can be monetized and what the relevant nodes and exchanges for the value creating and capture are. The concept of value extract is useful to gain focused view on what is actually relevant in the ecosystem for monetization. Each business in the ecosystem needs to have something that is beneficial for them from a business point of view. Value extracts can be single activities, automated processes, individuals, commercial organizations, non-profits, or even groups of organizations or networks with their respective value flows between their nodes.
Westerlund et al.'s business model framework heavily focuses on the value part of business models. Based on the concept of value design, those four value pillar described above come together in a single picture. Value design is an architecture mapping the founda-

tional structure of the ecosystem business model. It provides boundaries for an ecosystem and gives a pattern of operation.

10.5.3 Three-dimensional Collaborator Model Framework

The three-dimensional collaborator model presented by Chan [5] builds on a model described by Holler et al. [14]. The three dimensions of Chan's model are 'Who', 'Where', and 'Why'. On first sight, those dimensions seem similar to the nodes in Gassmann et al.'s 'Magic Triangle' proposition but the differences become evident once the meaning behind the three dimensions is investigated. The 'Who' describes the collaborating partners that build the value network. The 'Where' refers to sources of value co-creation and lastly the 'Why' describes how partners actually benefit from collaborating within the value network [5]. Table 10.1 shows the proposed framework. While those three dimensions are not explicitly visible, the individual components in the framework address them indirectly.

Table 10.1: Chan's three-dimensional collaborator model [5]

Collaborator	Inputs	Network	Service/ Processing/ Packaging	Content/ Information product	Benefits	Strategy	Tactic
ABC	C1 C2 C3						

Chan applied his framework to multiple case studies in order to explore the 'Who', 'Where' and 'Why' elements of his business model framework. Those case studies are thoroughly described in his paper *Internet of Things Business Models* [5]. Interesting are the columns 'Strategy' and 'Tactic'. Li et al. [20] proposed four IoT strategy categories which are adapted by Chan in the use cases for his three-dimensional business model framework. The strategies are 'get-ahead strategy in market', 'catch-up strategy in market', 'get-ahead strategy in technology', and 'catch-up strategy in technology'. Get-ahead strategies enable the firm to stay ahead of other competitors, giving them a first mover advantage. Catch-up strategies, on the other hand, are intended to follow and learn from the industrial leaders by operational efficiency and quality [5]. Besides strategy, the business model framework contains the 'Tactic' column. The six components of the 'Digitally Charged Products' pattern by Fleisch et al. [10] as presented in Section 10.4 are the proposed tactics to choose from.

10.6 Case Study Applying Three IoT Business Model Frameworks

In this section, a fictitious company is introduced in order to make it possible to compare the available business model frameworks and illustrate their respective strengths and weaknesses.

The case study company is called 'FlexSpace'. It offers an integrated beacon and a software solution for companies to make the usage of companies' office space more flexible by showing the available desk places in an office building to employees. This allows the customer companies to use office buildings more efficiently and thereby reduce the square meters per employee which directly saves money for the company. A beacon is a miniature, battery-powered radio transmitter, that usually employs Bluetooth Low Energy (BLE) technology to communicate with nearby devices. These beacons allow letting the

employees to check-in at their desk by using their smartphones. 'FlexSpace' offer the installation and maintenance of the described beacon infrastructure at customer sites as well as a white-label mobile application for their customers. The application displays the available desk places and enables the user to check-in at a specific desk using the closest beacon available. The visual appearance can be customized per customer to fit the customer company's corporate design.

10.6.1 Applying the Adapted Business Model Canvas for IoT

Based on the IoT-adapted BMC framework by Dijkman et al., a business model for 'FlexSpace' was created. The business model framework presented in Figure 10.5 was filled out and customized for the needs of 'FlexSpace'. The resulting business model is shown in Figure 10.7.
The 'Key Partners' block contains the producers of the beacon hardware. For this building block, it is crucial to have a 'launching partner'. This needs to be a big customer that finances the initial development. A suitable partner would be the University of Zurich (UZH) as it has many learning spaces, which could be managed using the beacon-based approach. Other partners are the beacon installation partner as well as a web hosting provider for the corporate website enabling account management. The 'Key Activities' are software development as well as traditional activities such as marketing, sales or customer contact. The 'Key Resources' needed are mainly employees with respective skill sets, the beacon hardware and the corresponding software. In the 'Cost Structure' block, the relevant costs for running the 'FlexSpace' business are listed. The important ones are development costs, salaries and marketing and sales. Dijkman et al. [8] rated 'Value Proposition' as the most important building block of a business model. For 'FlexSpace', the following value proposition were found important:

- The overview over all working spaces allows more convenient searching for free work places.
- Due to remote reservation possibilities, a higher comfort can be achieved.
- The available usage space data allows the analysis of work space usage for further optimizations.
- The customers of 'FlexSpace' are able to customize the beacons and the software to achieve high compliance with their corporate design.

The 'Customer Relationships' are maintained through the website as well as through a technical support hotline. The 'Channel' block describes how the customers will be reached. The primary channels are through the website and sales persons directly approaching customers. The important customers in 'Customer Segments' are companies or institutions with large numbers of working spaces. The revenue generation in the building block 'Revenue Streams' contains installation and customization fees of the beacons. The biggest revenue is generated through one-time sales of the beacons with the corresponding configuration software and recurring revenue through yearly licenses for the full Office Usage Software, including web-services and data analysis capabilities. It can be seen that a big part of the proposed framework from Dijkman et al. was used in the generation of the business model. Some proposed components were excluded, e.g. due to the chosen distribution channels (no partner stores). The resulting business model promises to be specific enough to be useful for the creation of the 'FlexSpace' business.

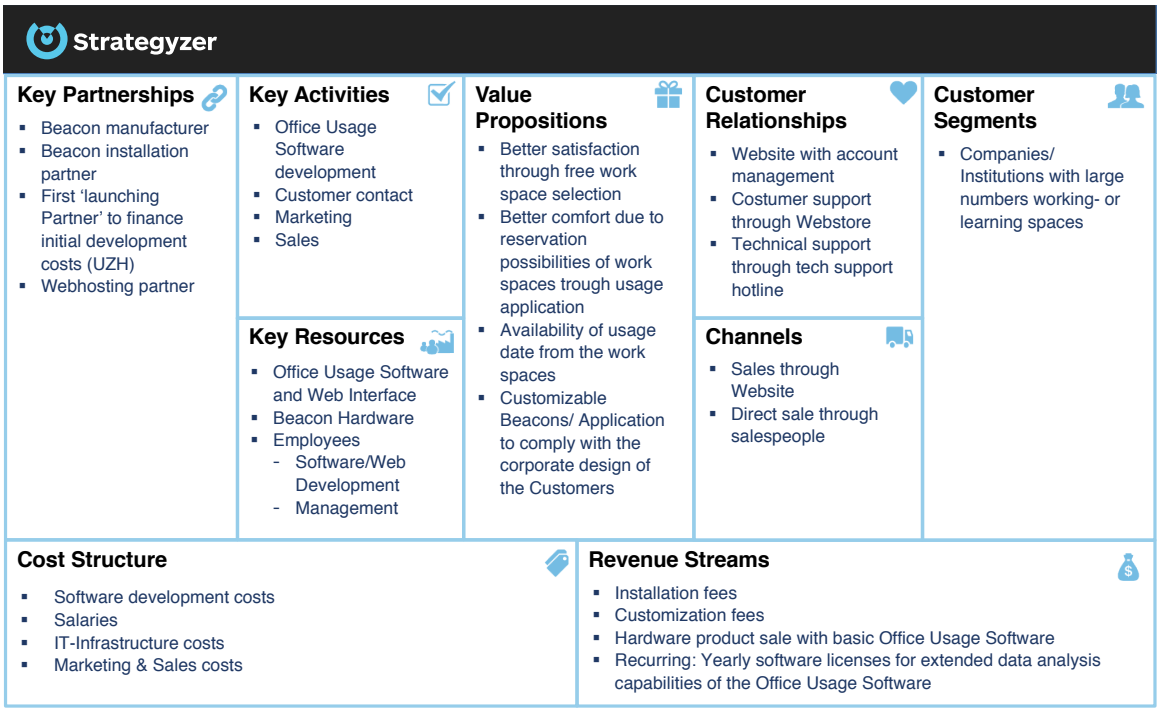


Figure 10.7: Business Model Canvas Framework applied to the 'FlexSpace' company

10.6.2 Applying the Ecosystem Business Model Framework

Westerlund et al. suggest "that managers need to shift their focus from 'the business model' of a firm to 'ecosystem business models'" [36]. The application of the framework proposed by them leads to the business model for FlexSpace in Figure 10.8. One of the

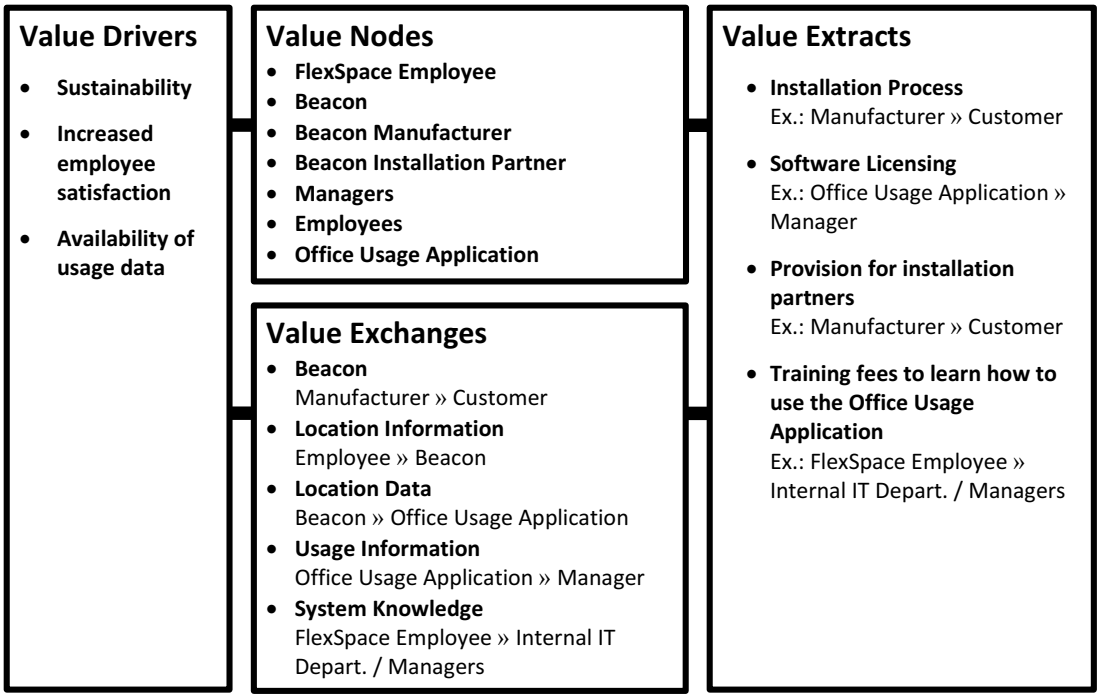


Figure 10.8: Ecosystem business model framework applied to 'FlexSpace'

'Value Drivers' behind the business idea of 'FlexSpace' is sustainability, which is realized in using less office space per employee thanks to the beacon solution. Also, employee satisfaction is expected to increase due to the fact, that employees can choose their own desk place flexibly after the introduction of the system. The availability of usage data is a starting point to turn the gathered data into value for the customer company.

The FlexSpace employees, beacons, beacon manufacturer, beacon installation partners, the customer's managers and employees, the application as well as the customer company's internal IT department make up the 'Value Nodes'. Note that there are two autonomous nodes, namely the beacons and the application. Some of the nodes are individuals (e.g., employees) and others are organizations (e.g., internal IT department).
'Value Exchanges' take place between all kinds of value nodes as indicated in Figure 10.8. Only a couple of these exchanges are relevant for 'Value Extract'. Namely the installation process, the licensing of software usage, the work of installation partners as well as training given to the customer are monetizable. It is, for example, hard to imagine a monetization strategy for the value exchange of location information between the employee and the beacon.

10.6.3 Applying the Three-dimensional Collaborator Model Framework

When Chan's framework is applied to the fictional company, the result shown in Table 10.2 can be gained. The emphasis on the mutual benefits for all collaborators becomes evident. Thanks to the beacons at each desk where the employees check-in, the target company now knows about the used capacity of workplaces. They can analyze which employees come on premise to work and which rather work from home. The software provided can measure changes in utilization over the year. Predictions for flexible allocations can be made to adjust the varying demand for workplaces over the year. Employees profit by being free to choose where they want to work and do not risk running into an occupied workplace again. FlexSpace's benefit is essentially the monetary reward.

Table 10.2: Chan's framework applied to the business idea of 'FlexSpace'

Collaborator	**Inputs**	**Network**	**Service/ Processing/ Packaging**	**Content/ Information product**	**Benefits**	**Strategy**	**Tactic**
FlexSpace	beacons/ software	IoT Network	algorithm for workplace usage predication and allocation	hardware network and software	monetary reward	catch-up strategy in technology	Digital lock-in
Target Company	work desks	company WLAN network	HR tool for workspace management	workspace usage and monitoring	workspace planning	get-ahead strategy in market	Product as Point of Sales
Target Company Employees	smartphone app			overview of free workspaces	job satisfaction and flexibility		

10.7 Comparison of IoT Business Model Frameworks

In this section, the results gained in Section 10.6 are evaluated and discussed. The three IoT-specific business model frameworks, which were applied to the case study company 'FlexSpace', are compared in terms of their usefulness when creating an easy to understand business model. After that, the changes in business models due to the emergence of IoT are discussed in general, based on our literature research presented in Section 10.2 and Section 10.3.

10.7.1 Evaluation of the Case Study

The business model framework from Dijkman et al. [8] is a template based on the BMC [32]. It contains the important building block types that should be considered in the

creation of an IoT business model. The template is easy to use and easily adaptable to custom needs whilst reminding to the networking perspective of IoT. The business model creation process does not depend on Dijkman's et al. framework, it could also be done using the classical BMC, albeit requiring more care to not unintentionally omit IoT-specific business model parts. With regards using it for 'FlexSpace', it can be said that the application was rather simple; the predefined business model parts of the framework reduced the complexity of the business model generation greatly.
Applying Westerlund et al.'s framework leads to the following observations: First, FlexSpace is not a typical ecosystem business idea. A typical example of a ecosystem business would be a multi-sided platform such as the Apple AppStore. This might be the reason, why the application of the framework was difficult. Also, the authors themselves, stated that their paper only proposes the four key value pillars, which can be used as a basis for the development of a practically applicable business model framework tool in the future. The application of Westerlund et al.'s framework showed clearly that more work is needed to turn this "plum pudding model" [36] into a practical tool for business model creation.
Chan's model proved itself already in many case studies in the original paper. Unsurprisingly, it also works really well for the case study and proved to be less abstract than the ecosystem framework by Westerlund et al.. It provides a comprehensive view, from how collaborators benefit from each other to what strategies and tactics are used. Same as the adapted BMC from Dijkman et al., it provides an easy and ready-to-use template to fill out. Chan's model goes with the ecosystem nature of IoT by addressing the different collaborators and their benefits, a perspective that is not found in the BMC.
Whatever approach businesses will choose in the future, the most important thing is to realize that IoT brings in perspectives, which are not explicitly addressed by traditional business model frameworks. Thus one is well advised to take action early and keep an open mind when it comes to business models for IoT. Some businesses may be better off sticking to the well-known BMC, while others may profit from adapting an ecosystem business model view. What businesses should definitely avoid is to fall for the well-known 'boiling frog syndrome' by not acting until it is too late. As long as there are still profits coming in, why take a risk and leave your comfort zone? Competition and environmental factors are always in movement. What works today may not work anymore in a few years. The 'boiling frog syndrome' states that gradually increasing problems go unnoticed or are not dealt with until it is too late to tackle them. A frog in a pot of water will not jump out of the water when it is slowly brought to boil. When the water gets too hot for him, his muscles are too weak to jump out. Likewise, a business that has diminishing returns year after year may not be concerned about the little loss between consecutive years. When they realize the devastating loss suffered over multiple years, there are no resources left to tackle the problem and they go bankrupt. That's why ignoring the changes IoT brings with itself may lead to a devastating result for companies unwilling to adapt.

10.7.2 Changes in Business Models due to the Emergence of IoT

Business models for IoT differ from traditional business models. This subsection takes a look at what changed due to the emergence of IoT.
Sticking to the BMC, Dijkman et al. [8] do not provide a new framework per se, but rather show the relative importance of the individual building blocks for IoT businesses. Unsurprisingly, the 'Value Propositions' building block has by far the highest relative importance compared to the other blocks. Generally speaking, an emphasis on value creation and capture as well as an emphasis on a more networked environment can be observed. Hui [15] states that in industries that are becoming connected, and differentiation, cost, and focus are no longer mutually exclusive. Rather than that, they can be mutually rein-

forcing in creating and capturing value.
In the ecosystem business model framework by Westerlund et al. [36], there is a dogmatic shift. The business model is no longer seen on the level of a single company like traditional business models do. Rather, the concept of value design is applied on an ecosystem level. Such a conception of business models in an ecosystem may be more suitable considering the nature of IoT. Networking, interconnections, interconnectivity and thus interdependence come with the territory when talking about IoT. IoT in itself is an ecosystem that can be scaled from only personal devices such as wearables to huge networks such as smart cities. In smart cities, there are many actors and stakeholders. Identifying ecosystems within a smart city and thus finding shared and private value drivers allow businesses to cooperate efficiently with each other in a win-win environment. Traditional business models lack an ecosystem view. With ecosystem business model frameworks, an important and inherent aspect to IoT can be addressed and monetized; something that other traditional business models fail to reach. Naturally such a business model becomes more complex and skilled managers are needed for a successful implementation. The focus on the firm level needs to be extended to a broader view. Connections to new industries have to be made which is not an easy task, but is needed for businesses to break into the emergent IoT market, prove themselves and offer value to their customers. Nevertheless, this ecosystem business model framework also shares some similarities with traditional business model frameworks. In its core, value design is a concept found in many business models. As can be seen in the 'Magic Triangle' by Gassmann [13], 'Value' is one of the nodes that make up the 'Magic Triangle' and is thus essential for a business model to work. Furthermore the pillars in the ecosystem business model framework can be categorized, compared and examined in the same way that the vertices in the 'Magic Triangle' or building blocks of the 'Business Model Canvas' can.
However, the ecosystem business model framework may be too abstract to be yet applied to businesses as can be seen in the case study. Westerlund et al. themselves state that they simply provided grounds for a novel tool for designing ecosystem business models which has to be further studied and developed. In contrast, Chan's three dimensional collaborator framework proves its usefulness in the case study. Similar to the ecosystem business model, it focuses on the multitude of actors who benefit from each other. Besides the emphasis on collaborators, concrete benefits for all participants are captured in the model and the new components 'Strategy' and 'Tactic' enrich the model. Chan's business model framework is specifically tailored for the needs of IoT companies.

10.8 Summary and Conclusion

This paper starts by presenting the essentials of the 'Internet of Things' paradigm and the concept of business models in general. After building a common ground, the classic business model frameworks were introduced. Based on literature research, the currently important business model frameworks for IoT businesses were then presented and described. The illustrated frameworks were compared based on an imaginary IoT model company. The application of the presented frameworks to the model company revealed that Dijkman et al.'s' framework provided the most complete and simple experience to create a new business model. Chan's collaborator framework was also ready-to-use, although it required more attention to get the 'complete picture' of the business model. Positively stood out that the ecosystem based-view on the constructed business model weighted the characteristics of network-oriented business models more than Dijkman et al.. Westerlund et al.'s value pillars framework was hard to apply to the model company. It focuses on the value-oriented part of business models and seemed overly abstract for a day-to-day usage scenario. A general comparison of the frameworks revealed an existing

shift from the classical single firm-based business model view to a more ecosystem-based structure of business models for IoT.
The literature on business models for IoT is relatively scarce, thus this paper is based on few self referencing papers, e.g. [18] and [8], and lacks a broader view, which reflects in this paper. Future research containing a large-scale field study could strengthen the current IoT business model frameworks, and help some frameworks to adapt and mature into business tools usable in real-world use cases.
To conclude, the only more generally usable framework is the 'Adapted Business Model Canvas for IoT' from Dijkman et al.. This framework finds the balances between the originally empty BMC and a completely predefined IoT business model solution. A deeper analysis of each block could lead to a more specific framework which would then in turn only be usable for a subsection of businesses.

Bibliography

[1] L. Achtenhagen, L. Melin, L. Naldi: *Dynamics of Business Models - Strategizing, Critical Capabilities and Activities for Sustained Value Creation*; Long Range Planning, Vol. 46 2013, 2013, pp 427-442, DOI: http://dx.doi.org/10.1016/j.lrp.2013.04.002, last visit: November 17, 2016.

[2] Amazon Echo: *Echo & Alexa, designed around your voice*: URL: https://www.amazon.com/dp/B00X4WHP5E, last visit: Dec. 30, 2016.

[3] L. Atzori, A. Iera, G. Morabito: *The internet of things: a survey*; Computer Networks, Vol. 54, 2010, pp 2787-2805, URL: http://www.science.smith.edu/~jcardell/Courses/EGR328/Readings/IoT%20Survey.pdf, last visit: Dec. 30, 2016.

[4] E. Bucherer, D. Uckelmann: *Business models for the Internet of Things*; In: Architecting the Internet of Things, Springer, Berlin-Heidelberg, Germany, 2014, URL: https://www.researchgate.net/profile/Ivan_Corredor/publication/259943140_Architecting_the_Internet_of_Things/links/0a85e52ea6e62681a1000000.pdf, last visit: Dec. 30, 2016.

[5] H.C.Y. Chan: *Internet of Things Business Models*; Journal of Service Science and Management, Vol. 8, 2015, pp 552-568, DOI: http://dx.doi.org/10.4236/jssm.2015.84056, last visit: Dec. 30, 2016.

[6] IEEE Standards Association: *IEEE-SA Internet of Things Ecosystem Study*; IEEE Standards Association, New York, 2015, URL: http://www.cisco.com/c/dam/en/us/solutions/collateral/industry-solutions/dlfe-670918525.pdf, last visit: Dec. 30, 2016.

[7] M.-H. Cho: *South Korea to invest $5b by 2020 in IoT and smart cars*; 2015, URL: http://www.zdnet.com/article/south-korea-to-invest-5b-by-2020-in-iot-and-smart-cars/, last visit: Dec. 30, 2016.

[8] R.M. Dijkman, B. Sprenkels, T.Peeters, and A. Janssen: *Business Models for the Internet of Things*; International Journal of Information Management, Vol 35, 2015, pp 672-678, DOI: http://dx.doi.org/10.1016/j.ijinfomgt.2015.07.008, last visit: Dec. 30, 2016.

[9] P. F. Fan, G. Z. Zhou: *Analysis of the business model innovation of the technology of internet of things in postal logistics*; 18Th IEEE International Conference on Industrial Engineering and Engineering Management(IE&EM 2011), Ghangchun, China, 2011, pp 532-536, URL: http://ieeexplore.ieee.org/document/6035215/, last visit: Dec. 30, 2016.

[10] E. Fleisch, M. Winberger, F. Wortmann: *Business Models and the Internet of Things*: Bosch Internet of Things & Services Lab, pp 1-19, August 2014, URL: http://www.iot-lab.ch/?page_id=10543, last visit: Dec. 30, 2016.

[11] Gartner: *4.9 Billion Connected "Things" Will Be in Use in 2015*; 2014, URL: http://www.gartner.com/newsroom/id/2905717, last visit: Dec. 30, 2016.

[12] O. Gassmann, K. Frankenberger, and M. Csik: *Revolutionizing the business model.* in O. Gassmann and F. Schweitzer (Eds.), Management of the Fuzzy Front End of Innovation. Springer, New York, USA, 2014, pp. 89-97, URL: http://link.springer.com/chapter/10.1007%2F978-3-319-01056-4_7, last visit: Dec. 30, 2016.

[13] Gassmann et al.: *Geschaeftsmodelle entwickeln: 55 innovative Konzepte mit dem St. Galler Business Model Navigator*; Carl Hanser Verlag GmbH Co KG, 2013, ISBN 978-3446435674

[14] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, S. Karnouskos, D. Boyle: *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*; Elsevier, 2014, ISBN 978-0-12-407684-6

[15] G. Hui: *How the Internet of Things Changes Business Models*; Harvard Business Review, July 24, 2014, URL: https://hbr.org/2014/07/how-the-internet-of-things-changes-business-models, last visit: Dec. 30, 2016.

[16] IFTTT: *Do more with the services youÂ love*; URL: https://ifttt.com/, last visit: Dec. 30, 2016.

[17] ITU: *New ITU standards define the internet of things and provide the blueprints for its development*; ITU-T recommendation Y-2060, URL: http://www.itu.int/ITU-T/newslog/New+ITU+Standards+Define+The+Internet+Of+Things+And+Provide+The+Blueprints+For+Its+Development.aspx, last visit: Dec. 30, 2016.

[18] Jaehyeon Ju, Mi-Seon Kim and Jae-Hyeon Ahn: *Prototyping Business Models for IoT Service*, Procedia Computer Science, Vol. 91, 2016, pp 882-890, URL: http://www.sciencedirect.com/science/article/pii/S1877050916312911, last visit: Dec. 30, 2016.

[19] Kickstarter: *Our mission is to help brin creative projects to live*; URL: https://www.kickstarter.com/, last visit: Dec. 30, 2016.

[20] Y. Li, M.J. Hou, H. Liu, Y. Liu: *Towards a Theoretical Framework of Strategic Decision, Supporting Capability and Information Sharing under the Context of Internet of Things*; Information Technology and Management, Vol. 13, 2012, pp 205-216, DOI: http://dx.doi.org/10.1007/s10799-012-0121-1, last visit: Dec. 30, 2016.

[21] H. Li, Z. Z. Xu: *Research on business model of Internet of Things based on MOP*; International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012), Proceedings, pp 1131-1138, Springer, Berlin Heidelberg, Germany, 2013, ISBN 978-3-642-38444-8

[22] LiFi Labs, Inc.: *LIFX*; 2016 URL: http://www.lifx.com/, last visit: Dec. 30, 2016.

[23] L. Liu, W. Jia: *Business model for drug supply chain based on the internet of things*; 2nd IEEE InternationalConference on Network Infrastructure and Digital Content, 2010, pp 982-986, URL: http://ieeexplore.ieee.org/document/5657943/?reload=true, last visit: Dec. 30, 2016.

[24] B. Miller: *Obama Places $160 Million Bet on Smart Cities, Internet of Things*; 2015, URL: http://www.govtech.com/Obama-Places-160-Million-Bet-on-Smart-Cities-Internet-of-Things.html, last visit: Dec. 30, 2016.

[25] M. Morris, M. Schindehutte, J. Allen: *The Entrepreneur's Business Model: Toward a Unified Perspective*; Journal of Business Research, Vol. 58, 2005, pp 726-735, DOI: http://dx.doi.org/10.1016/j.jbusres.2003.11.001, last visit: Dec. 30, 2016.

[26] A. Osterwalder, Y. Pigneur, C. L. Tucci: *Clarifying Business Models: Origins, Present and Future of the Concept*; Communications of the Association for Informa-

tion Science, Vol. 16, 2015, pp 1-25, URL: http://aisel.aisnet.org/cais/vol16/iss1/1, last visit: Dec. 30, 2016.

[27] A. Osterwalder, Y. Pigneur, A. Smith: *Business Model Generation*; John Wiley and Sons, 2010.

[28] A. Osterwalder, Y. Pigneur: *Business model generation: a handbook for visionaries, game changers, and challengers*; John Wiley & Sons, Hoboken, NJ, USA, 2010.

[29] M.E. Porter, J.E. Heppelmann: *How smart, connected products are transforming competition*; Harvard Business Review, Vol. 92, 2014, pp 11-64

[30] B. Rossi: *How the Internet of Things is Changing Business Models*; Information Age - Insight and analysis for IOT leaders, May 4, 2016, URL: http://www.information-age.com/it-management/strategy-and-innovation/123461371/how-internet-things-changes-business-models, last visit: Dec. 30, 2016.

[31] L. Schweizer: *Concept and Evolution of Business Models*; Journal of General Management, Vol. 31, 2005, pp 37-56.

[32] Strategyzer AG: *The Business Model Canvas*; URL: https://strategyzer.com/canvas/business-model-canvas, last visit: Dec. 30, 2016.

[33] Y. Sun, H. Yan, C. Lu, R. Bie, P. Thomas: *A holistic approach to visualizing business models for the internet of things*. Communications in Mobile Computing, Vol. 1, 2012, pp 1-7, DOI: http://link.springer.com/article/10.1186/2192-1121-1-4, last visit: Dec. 30, 2016.

[34] A. Tilley: *Google Acquires Smart Thermostat Maker Nest For $3.2 Billion*; 2014, URL: http://www.forbes.com/sites/aarontilley/2014/01/13/google-acquires-nest-for-3-2-billion/#7049b3181416, last visit: Dec. 30, 2016.

[35] A. Tilley: *Samsung Acquires SmartThings, A Fast-Growing Home Automation Startup*; 2014, URL: http://www.forbes.com/sites/aarontilley/2014/08/14/samsung-smartthings-acquisition-2/#25b90c197965, last visit: Dec. 30, 2016.

[36] M. Westerlund, S. Leminen and M. Rajahonka: *Designing Business Models for the Internet of Things*; Technology Innovation Management Review, July 2014, URL: http://timreview.ca/article/807, last visit: Dec. 30, 2016.

[37] F. Wortmann and K. Fluechter: *Internet of things. Business & Information Systems Engineering*; Vol. 57, 2015, pp 221-224, http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1338&context=bise, last visit: Dec. 30, 2016.

[38] C. Zott, R. Amit, and L. Massa: *The business model: recent developments and future research*; Journal of management, Vol. 37, 2011, pp 1019-1042, URL: http://journals.sagepub.com/doi/pdf/10.1177/0149206311406265, last visit: Dec. 30, 2016.

Bibliography

[1] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1617–1634, 2014.

[2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, 2015.

[3] O. N. Fundation, "Software-defined networking: The new norm for networks," *ONF White Paper*, 2012.

[4] J. Burke, "software-defined networking (sdn)," 2013. [Online]. Available: http://searchsdn.techtarget.com/definition/software-defined-networking-SDN

[5] ——, "What is SDN? The answer now includes automation and virtualization," 2015. [Online]. Available: http://searchsdn.techtarget.com/tip/What-is-SDN-The-answer-now-includes-automation-and-virtualization

[6] E. Banks, "Sdn basics: Understanding centralized control and programmability," 2014. [Online]. Available: http://searchsdn.techtarget.com/tip/SDN-basics-Understanding-centralized-control-and-programmability

[7] M. McNickle, "Five must-know open source sdn controllers," 2014. [Online]. Available: http://searchsdn.techtarget.com/news/2240225732/Five-must-know-open-source-SDN-controllers

[8] N. Sharma, "Eight big benefits of software-defined networking," 2015. [Online]. Available: http://www.serverwatch.com/server-tutorials/eight-big-benefits-of-software-defined-networking.html

[9] B. Fraser, D. Lake, C. Systems, J. Finnegan, N. Viljoen, and S. O. E. N. Etworking, "Are we ready for sdn ? implementation challenges for software-defined networks," no. July, pp. 36–43, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6553676/citations

[10] A. Warfield, "Data, storage, and sdn: An application example." [Online]. Available: http://www.cohodata.com/blog/2014/01/22/data-storage-and-sdn-an-application-example/

[11] M. Rouse, "data plane (dp)," -. [Online]. Available: http://searchsdn.techtarget.com/definition/data-plane-DP

[12] ——, "control plane (CP)," 2016. [Online]. Available: http://searchsdn.techtarget.com/definition/control-plane-CP

[13] SDXcentral, "What are SDN Northbound APIs?" 2016. [Online]. Available: https://www.sdxcentral.com/sdn/definitions/north-bound-interfaces-api/

[14] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, 2015.

[15] M. Rouse, "Network Functions Virtualization (NFV)," 2016. [Online]. Available: http://searchsdn.techtarget.com/definition/network-functions-virtualization-NFV

[16] A. Lemke, "How to manage security in nfv environments," online, 2014. [Online]. Available: https://insight.nokia.com/how-manage-security-nfv-environments

[17] C. Matsumoto, "Nfv performance should be a bigger issue." [Online]. Available: https://www.sdxcentral.com/articles/news/nfv-performance-bigger-issue/2015/01/

[18] G. ETSI, "Network functions virtualisation (nfv): Architectural framework," *ETSI GS NFV*, vol. 2, no. 2, p. V1, 2013.

[19] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[20] N. Omnes, M. Bouillon, G. Fromentoux, and O. Le Grand, "A programmable and virtualized network & it infrastructure for the internet of things: How can nfv & sdn help for facing the upcoming challenges," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015, pp. 64–69.

[21] S. Bagchi, "Unlocking the potential of sdn and nfv." [Online]. Available: http://blogs.windriver.com/wind_river_blog/2016/08/unlocking-the-potential-of-sdn-and-nfv.html

[22] C. Mathas, "What's next for telco sdn/nfv? fast growth, slower deployment." [Online]. Available: https://itu4u.wordpress.com/2016/06/28/whats-next-for-telco-sdnnfv-fast-growth-slower-deployment/

[23] F. A. Khan *et al.*, "Virtualized epc: Unleashing the potential of nfv and sdn," in *25th European Regional ITS Conference, Brussels 2014*, no. 101426. International Telecommunications Society (ITS), 2014.

[24] M. Dano, "How verizon, at&t, t-mobile, sprint and more stacked up in q2 2016: The top 7 carriers." [Online]. Available: http://www.fiercewireless.com/wireless/how-verizon-at-t-t-mobile-sprint-and-more-stacked-up-q2-2016-top-7-carriers

[25] J. Doherty, *SDN and NFV Simplified: A Visual Guide to Understanding Software Defined Networks and Network Function Virtualization*. Addison-Wesley Professional, 2016.

[26] A. D. Little, *Reshaping the future with NFV and SDN*. Bell Labs Alcatel-Lucent, May 2015.