JOHANNES KEPLER
UNIVERSITY LINZ | JKU

TNF

Faculty of Engineering
and Natural Sciences

# Seminar in Cloud Computing
# IaaS Systems


Submitted by:
Markus Hiesmair
Jürgen Ratzenböck
Christoph Stenglein

Created at:
Institute for Intelligent Information Systems

Graded by:
Univ.-Prof. Dr. Roland Wagner


Linz, January 12, 2016

# Abstract

Taking the right decision when buying infrastructure to deploy your huge and complex software system can often be difficult nowadays. Infrastructure as a Service (IaaS) has been production ready for some years now and gains high popularity among productive and successful businesses. Instead of hiring dedicated hardware at your hosting provider of choice with your fixed capacity at the preferred destination, IaaS tries to tackle and solve common problems of the traditional approach. IaaS offers virtualized infrastructure without any upfront capacity decisions to provide a highly available, scaleable and flexible system for your software. This paper introduces IaaS as moving infrastructure to the cloud and illustrates it's opportunities and benefits. The paper outlines the technical details and explains how IaaS works under the hood, how it can be managed, integrated and monitored. We look at big internet companies behind the technology and their own infrastructure model. It will be discussed why customers can save money by buying infrastructure in the cloud and paying only for the capacity the really use and at the same time increase their revenue due to the simple management of instances.

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Overview on Cloud computing

Nowadays software isn't just installed on an arbitrary computer for a specific user who can fulfil his given requirements by solving a task with it. Quite the contrary is the case as the significance of software has increased dramatically throughout any kind of business sector. The demand on software products these days is immense and therefore also the complexity and variety has experienced a huge growth over the last decade. Many years ago the Internet built up the fundament of accessing and sharing information worldwide and today applications and services, relying on complex and huge software ecosystems, give people around the globe the opportunity to use them any time and anywhere they want to satisfy their needs. To make this work this obviously needs a lot of resources accessible in the global network.

Here the famous and hyped term "Cloud computing", which describes the process of moving application and services to the internet (due to the schematic metaphor also denotes as "cloud"), comes into play. [5] In such intensive businesses with rare resources as we have it nowadays people have to concentrate on their specific tasks to be as productive, competitive and flexible as possible. Cloud computing supports this by providing a pool of resources allowing for sharing and scalable deployment of services, as needed, from almost any location, and for which the customer can be billed based on actual usage. [5]

How these resources are provided and shared depends on the specific requirements and can vary. Due to the common patterns of usages some different cloud types describing the strategy have established over time. [5]

- **Private Cloud**: The sharing of resources stays in-house and a specific organization is responsible for operating and maintaining the cloud infrastructure.

- **Community Cloud**: Several organizations having a common interest operate and maintain the shared cloud infrastructure. For the participating organizations such a solution can be very cheap if they agree on the community model.

- **Public Cloud**: An organization renting the cloud infrastructure from a specific provider who is responsible for it. The infrastructure is publicly available on a commerical basis.

- **Hybrid Cloud**: This is a mixture of the other existing types which can be tailored based on the concrete requirements for optimizing productivity. This can be for instance used if some data should be necessarily kept in-house and the rest could be outsourced in a Public Cloud.

Over the years different service models depending on the type of the provided resource have been established. Bascially they can be divided into three different types organized in a cloud computing stack with increasing abstraction level bottom-up.

**IaaS** basically means providing a shared pool of compute, storage and networking resources to end-users on a self-service basis. [6] This should help the end-users avoiding additional costs by buying dedicated hardware and setting up the instances to run their applications. They can easily manage and control the systems, in terms of operating system, network connectivity and storage and applications running on these instances but do not have to care about controlling and maintaining the cloud infrastructure. [5]

**PaaS** as the name already indicates provides the whole platform ”out-of-the-box” to the end-user. This includes things like the operating system or network

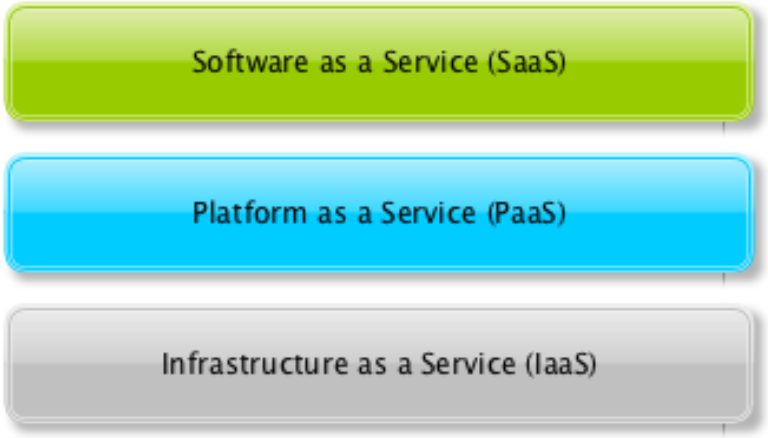


Figure 1.1: Stack of service models

connectivity which are completely managed by the provider. The user only has to deploy her applications to the cloud. [5]

**SaaS** abstracts the platform and infrastructure and serves the software living in the cloud as a usable service to the end-user. [5] This gives users instant access to such software without any special requirements such as downloading or installing and enables cross-platform as well as cross-device possibility.

### 1.1.1 Differences between IaaS, PaaS and SaaS

What distinguishes IaaS from PaaS or SaaS are the main areas of operation: IaaS operates completely as Service for computer infrastructure like hardware, storage, servers or network components. Compared to PaaS and SaaS, the users are responsible for managing the applications, data and OSes. The provider still manages virtualization, storage and hardware.

PaaS provides a platform used for applications and development. The user manages his application and the deployment of these. The provider manages OSes, virtualization, storage and the PaaS software. PaaS allows the user to create application with the PaaS software as middleware. This reduces the amount of work for managing an developing environment.

SaaS is the largest cloud market and uses the web to deliver applications which are managed by the provider and used by the clients. Most of these applications can

be run from a web browser, which makes it unnecessary to install the applications on indvidual computers. The user doesn't have to care about anything here; he just uses the application. The provider manages the application, runtime, data, OSes, hardware etc. [12]

# Chapter 2

# Advantages and Disadvantages of Cloud Computing Systems

The classical approach of deploying software, which was the case before the invention of cloud computing, is called ”Dedicated Hardware”. Using this approach companies buy hardware on their own which is dedicated only for the intended software, which should run on it. To provide a good level of service, usually companies buy hardware which can handle worst case scenarios and load peaks [8].

The problem of this approach is, that if the system runs on average work load, the full hardware capacities are not used and therefore resources are wasted (e.g. CPU cycles, storage space or RAM). If there are peaks in the work load this can result in a huge waste of resources or in a poor service of the software, if the hardware is not designed for such high peaks. Figure 2.1 shows some of these problems [8].

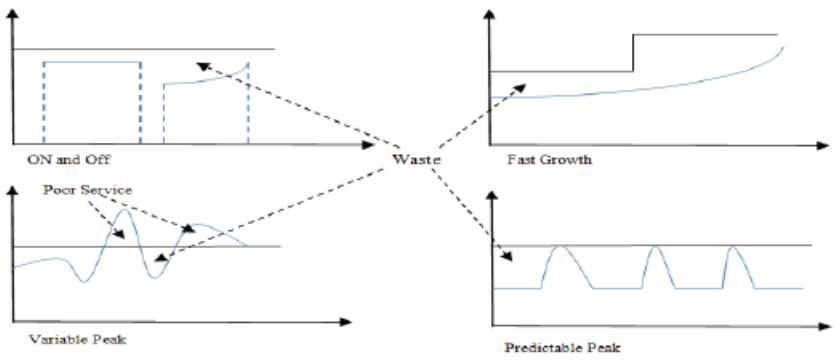


Figure 2.1: Dedicated Hardware Model - Utilization Waste [8]

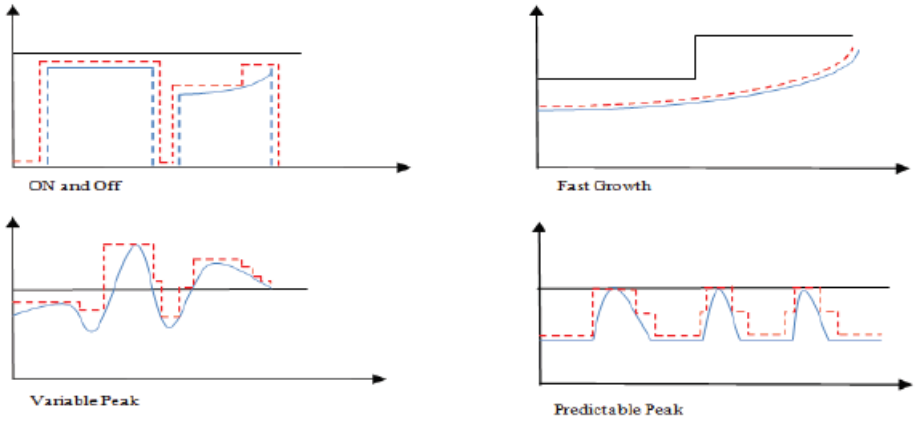


Figure 2.2: Cloud Computing Model - Dynamic Resources [8]

In cloud computing this waste of resources can be prevented by dynamically assigning resources when needed. For example, if the cloud computing system identifies that there is an overload, it can just add another virtual machine which can also handle requests. Later if there is less workload this VM can be shutdown and the available resources can be used for other services. Figure 2.2 shows the dynamic resource allocation of cloud computing. As you can see the changing resource need is handled in real time in cloud computing environments and therefore the resource utilization is being increased [8].

Another advantage is, that the cloud computing user does not have to care about backing up data, load balancing and resilience planning and still has highly scaleable and highly available applications. Detailed information about the technical details are described in chapter 3.

Probably the only disadvantage of hosting applications in the cloud is, that you have to fully trust the cloud computing provider, because every system acts like a black box. This means that the user does not know where the applications and data are stored and how they are protected against unauthorized access. The data can be distributed in many data centers in many countries to enhance scalability and availability. For high sensitive data, (e.g. online banking systems) cloud computing with a public provider is obviously not a good choice.

# Chapter 3

# Technical Details

## 3.1 Virtualization

The most important concept of cloud computing and therefore also for IaaS systems is virtualization. Nowadays every cloud computing system is virtualized. This means that all the user's cloud computing applications run in virtual machines. In the case of IaaS systems the user buys a VM and can install whatever system he likes on it [1].

Virtualization brings many advantages. A lot of crucial concepts of cloud computing are a lot easier with virtualized systems. On example is load balancing: Virtual machines can be copied and can be migrated on other physical machines without any problems - Figure 3.1 illustrates this graphically. The following section will give you a greater insight into load balancing strategies [1].

## 3.2 Load Balancing

As mentioned in the previous section, the applications on cloud computing systems run in virtual machines, and these VMs run in physical machines. Furthermore, there is a huge variation on the load (or resource needs) on the applications, therefore if there run too many applications (or VMs) on one physical machine it may get overloaded. This is why one needs load balancing. Load balancing should avoid that the physical machines have to handle more resources than they

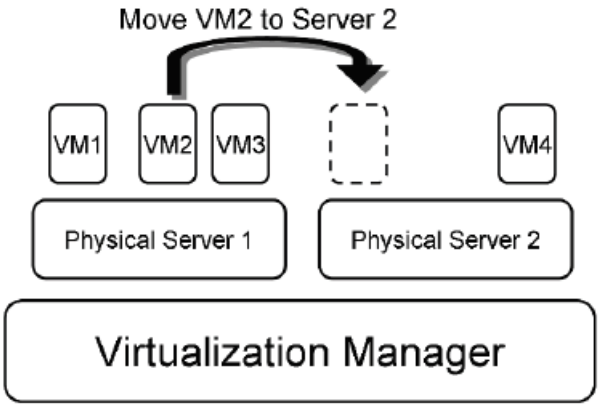


Figure 3.1: Virtualization and Live VM Migration in Cloud Computing Systems [1]

can offer. These resources can be CPU, RAM or storage space. If there run too many VMs on one physical machine this can mean a tremendous slow down of the VMs running on it. But slow VMs and applications are not the only problems, that bad load balancing would cause. If a application does not have sufficient resources, often the Service Level Agreement (SLA) is violated. A SLA is the agreement between customer and provider that guarantees certain parameters like performance and availability. A violation can mean that the cloud computing provider may have to give discount because of the violation or the customer decides to change to another cloud computing provider. As you can see good load balancing is crucial in the field of cloud computing [3].

### 3.2.1 Load Balancing Algorithms

There are many approaches on how load balancing algorithms should work, but they all have something in common: the input and the output. Every algorithm has to watch over the states of the physical and virtual machines and has to decide if a VM gets moved from. Furthermore, it has to decide which VM has to be migrated to which physical machine [3].

Many common load balancing algorithms are based on the current states of the PMs and VMs. They therefore are called ”reactive” algorithms. This means that when the resource utilization on a certain PM reaches a certain threshold, a VM gets migrated to another PM. This algorithm is rather easy to implement, as it

only has to watch the current resource utilization at the physical machines, but the disadvantages are that it only considers the current state of the system and that when a the threshold is reached most often, an imbalance situation is yet the case. Furthermore, it cannot guarantee a long-term balance situation, as it only acts on the parameters known at that point of time [1, 3].

Other algorithms are based on a "proactive" approach. In these algorithms the physical machines try to predict the resource demand of the VMs running on them and if in the near future there would be a overload they migrate a VM to another physical machine which predicts a lower resource utilization. This has the advantage that if the algorithm works as desired and the estimates on the resource demands of the VMs are approximately correct, there won't be any more overloads, because the algorithm would predict correct and migrate the VMs before the physical machine is overloaded. Usually the proactive approaches give better results than the earlier discussed reactive approach. But, however, there are also disadvantages. The first one is that the physical machine does under usual circumstances not know, which VM should be migrated to another PM. Furthermore, in the long run this approach also does not create a balanced state, because it only predicts a certain time span in the future. Finally, there have to be made a lot of calculations to predict the resource needs of the VMs (this is usually done by a Markov model), especially when there are a lot of VMs per physical machine. This often creates a big load just for the load balancing algorithm [2, 3].

## 3.3 Resilience Planning

As customers in nowadays IT world expect applications to be accessible anytime, anywhere from any device an "always-on, always-available" service is pervasive. As with every technical product also parts of an cloud infrastructure can experience several failures and to provide high availability it is essential to consider resilience when operating services in the cloud [17]. Data and it's volume to be managed is getting more complex and made high-availability solutions and disaster recovery strategies mandatory.

An important point is to address the evolvements and advances brought by the cloud in the resilience planning process. In contrast to dedicated hardware or

single physical servers, IaaS provides heavy virtualization of physical compute resources which can be put together in a single software-based virtual network. Due to this reasons the possibility for failures, especially on the software side, has increased and it's clear that errors will occur [17]. There are two metrics driving resilience:

- Recovery Time Objective (RTO): This is basically the time it takes to restore a service after it has experienced a failure.

- Recover Point Objective (RPO): This is an indicator for the data los after a failure.

To make an IaaS cloud infrastructure resilient geographic distribution of data centres connected via WAN plays an important role. On the one hand latency can be decreased due to the smaller distance between an edge location providing the service and the end-user requesting it and on the other hand geographic spreading adds redundancy because the same service can be provided from another data centre possibly being thousands of kilometres away from the affected one. During resilience planning you first have to identify which risks you want to consider and how to set the constraints for your service level agreement. These results are the basis for modeling the infrastructure distribution. Important tasks are investigating the underlying network and connections to identify so called shared risk groups (SRGs) which suffer in case of a specific failure. With this knowledge you can decide which mechanisms should be taken. Before any recovery can be started, failures have to be detected which by periodic heartbeats or network alarms of infrastructure components. After that a healthy component starts running the affected service and traffic must be redirected via networking reconfiguration subsequently to prevent further requests to the unhealthy instance [7]. How these backup strategies can look like will be explained in more detail in the following section.

## 3.4 Backup Strategies

A backup is a copy of important files and is usually made to reduce the chance of data-loss.

Backups don't have a high priority for the average user and most of them don't even back up their data [13] but the loss of data is very costly and time-intensive if you don't take precautions in form of a backup.

One Backup strategy would be the „3-2-1 rule" which is widely known. It describes a basic and simple strategy and avoids a single point of failure. The following principles are the core strategy:

- Have at least three copies of data
- In two different formats
- with one of those copies off-site

The three copies are including the one you already use, so you would have 2 backups, stored on 2 different media types (e.g. hard drive and DVD) in a different format and one of these have to be off-site. That means it should be stored somewhere else than the other copy (e.g. Out of the house).

This prevents single point of failure: When one hard drive is damaged, you have a second copy. If the format is corrupted, you have another backup in a different format. If something happens to the location, you have one backup off-site, since one event can't destroy both copies. This strategy doesn't prevent data-loss, but it prevents most of the cases [20]. IaaS comes here into play in multiple parts: You can store backups online to have them off-site. You don't have to buy extra hardware and store them somewhere else. You can store your backups on different cloud providers to prevent data-loss with one. Many cloud providers offer software to automatically backup your data to their cloud so there is less work for the user in that regard. In addition, cloud providers have big datacenters and have measures against data-loss, so it is safer to store one copy there [10].

## 3.5 Monitoring

Monitoring is an essential part for deployed systems for both, dedicated hardware structures and cloud computing systems. It enables to discover and analyze failure, bad configuration, performance bottle necks and many other issues that are

important parts of software maintenance. Cloud computing systems are made to be highly scaleable and elastic. But these two characteristics make it hard to monitor cloud computing systems as monitoring of high scaling systems requires to collect, store and analyze a lot of data in real time, which can be computationally highly expensive. Furthermore, the high elasticity shows an extra challenge, as the whole environment can change in a very short period of time in modern clouds [21].

### 3.5.1 Common Architectures for Distributed Systems

The most important architectures that are typical for monitoring distributed systems are:

- Flat Pull Model: A central server polls all the machines he should monitor according to a schedule when machines join and leave.
- Hierarchical Pull Model: Monitoring servers poll a subset of all the servers which should be monitored. A central server then polls the monitoring servers.
- Hierarchical Push Model: The server themselves push information to a monitoring server which is assigned to them and the monitoring servers collect that information from their machines and push it to a central server [21].

However, the just mentioned models for monitoring are modeled for grid and cluster computing. They are not always optimal for cloud computing systems as virtual machines can dynamically be added and removed and this creates the need for other approaches [21, 11].

### 3.5.2 Architectures modeled for Cloud Computing

In 2010, Huang [11] proposed a monitoring system he called ”Push and Pull”, which should combine the advantages of the push and pull. In current pull systems, the consumers (monitoring servers) pull information from the producers

(virtual machines which should be monitored) in a certain interval. This approach is efficient but lacks in consistency. If the interval is too big, data is lost, but if the interval is too small it is inefficient as there is too much network traffic. In push systems on the other hand, producers tell the consumers changes whenever the changes are greater than a threshold. This can produce good performance depending on the threshold and all data changes are monitored, but if the threshold is too small, too much information is transmitted and this leads again to inefficiency. The Push and Pull model Huang suggests, should use both approaches in a mixed way depending on the situation. Therefore, Huang introduces a value called User Tolerant Degree (UTD). This value indicates if high accuracy is a core requirement for the user or if the user tolerates minor inaccuracy. In Push and Pull both approaches are used simultenously in different degrees. For example, if the UTD is small, the user won't tolerate inaccuracy and the push model will dominate over the pull model. This means that according to the UTD, the producer will push changes larger than a relatively small threshold to the consumer, but at the same time the consumer will pull with a long interval to be sure not to miss changes. But if the UTD is large (the user will tolerance inaccuracy), the pull model will dominate. The consumer will have a short interval in which he pulls information from the producer and the producer only pushes changes if they are bigger than a large threshold.

Another approach was proposed from Ward [21] in 2014. He called his system VARANUS and it is built up on a layered probabilistic multicast or gossip protocol. By dividing up computational complexity over the system, gossip protocols show great performance in large scale networks. Within the layers the peers communicate via push and pull with other peers which are nearby. The distance between peers is determined by measuring the round trip time between the peers. In VARANUS the communication is different from layer to layer. In the lower layers, there is a great bandwidth and therefore the exchange rate is high and the information is highly consistent, whereas in the higher layers, the information is sent in lower intervals, because else there would be a too high network traffic. Figure 3.2 shows the architecture of the VARANUS.

Both architectures show significant improved results in terms of performance, latency and scalability in large scale deployment systems than the traditional systems for cluster and grid computing (discussed in section 3.5.1).

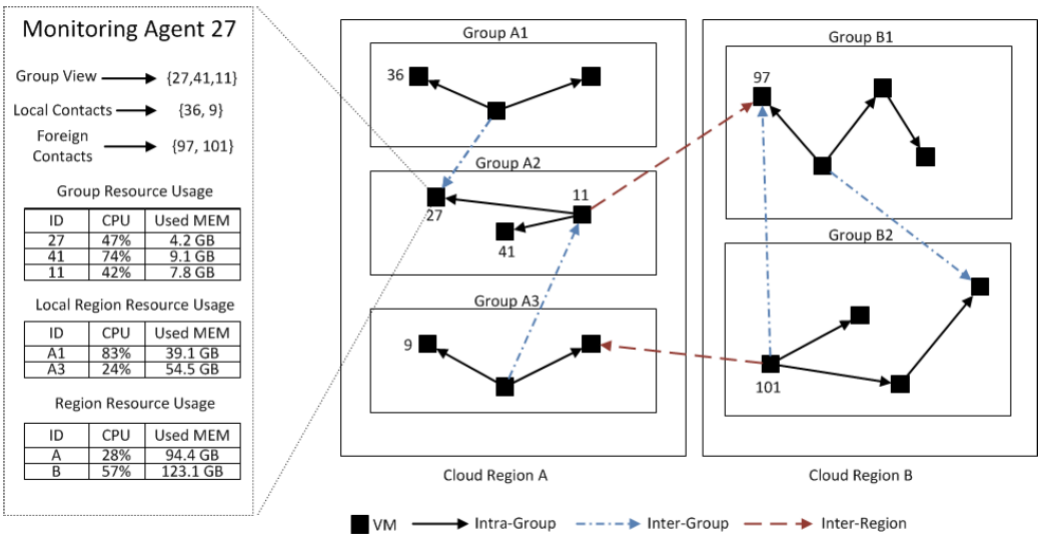


Figure 3.2: Architecture of the VARANUS monitoring system [21]

# Chapter 4

# Security

Security is very important for sensitive data, especially when hosted somewhere on the internet. This requires strong security but also practicable authentication. Common methods include:

- Simple authentication
- Multi factor authentication
- Certificates

In this chapter, we will first take a look at different authentication techniques. After that, we will compare the advantages and disadvantages in security in the cloud.

## 4.1 Authentication

### 4.1.1 Simple authentication

This method only requires a username and password and is really simple and widely used. However, it may be possible to find out your password, and if that happens, you don't have much chance to stop the intruder and lose access at least for some time.

### 4.1.2 Multi Factor Authentication

Multi factor authentication requires another factor of authentication in addition to your password. This could e.g. be a trusted device. Since many people use a smartphone these days, this could be a potential second factor to use for MFA. This authentication can be either a call or a SMS sent to your phone with a code you have to enter in addition to your password. This prevents intruders to gain access when they find out your password and is much more secure than using just a password. Another device would be a TOTP (Time-base One-Time Password) like a Yubikey. These generate passwords which only can be used once and are only valid for a limited time [19].

### 4.1.3 Certificates

Another way of authentication would be through certificates. These use asymetric cryptography where your opponent can use your public key to verify your identity and vice versa. These certificates can either be created by yourself (self-signed) or by a Certificate Authority. The certificate is then uploaded to the webserver and configure it.

You can also create certificates for the user and specify if the connection is allowed by checking his identity through his certificate [4].

## 4.2 Advantages and Disadvantages in Security

### 4.2.1 Advantages

#### 4.2.1.1 Easier Security Management and Maintenance

One advantage of using IaaS is that the Security management is much more professional since the providers Job is only to secure the datacenter. In addition, since it has a much bigger scale than hosting your own server, the provider uses much more redundancy and physical security than a „normal“ datacenter. This

provides a better security against data-loss both on the logical(read/write error on the hard drive) and the physical(fire, intruders) level with a cheaper price overall [18].

#### 4.2.1.2 Better trained personal

Many industries require certitifcations in information security and or privacy. These can be expensive for smaller companies who just manage a small amount of servers. Cloud vendors however often get their employees and their company certified to better train their employees and ensures more security overall. In addition, it demonstrates their knowledge and builds up trust to their customers.

### 4.2.2 Issues

IaaS does provide nice advantages compared to using your own datacenter but it also has some issues, which will be discussed in this chapter.

#### 4.2.2.1 Trust

One issue with IaaS could be trust. Any provider, be it IaaS, PaaS or SaaS, could have access to sensitive data since it is stored on their hardware. Insider access can be done by current or even former employees if security policies aren't enforced all the time. This can include regular password switching, requiring Multi factor authentication for employees etc. Another issue can be the monitoring of these services. If you migrate to cloud services, the responsibility for securing these systems where the customers data is operating on, goes to the provider. This can cause a big risk for the company if the cloud provider does provide sufficient security on their systems.

#### 4.2.2.2 Infrastructure

A issue, which comes together with trust, is the infrastructure. Having your sensitive data outsourced on a big cloud provider makes you a bigger target.

They may not target a single company which is using their services specifically, but since they target the provider, they may also get access to a company's sensitive data. The main focus of protection is mostly on the data where the customers are working on but there can be other valuable data to get stolen as well. For example payment information or user accounts of the services. But not only server-side protection is important, client-side protection is important as well. The Server is accessible on the internet and not avaliable on the intraweb, so you have to have a more secure connection to your server [15].

#### 4.2.2.3 Avaliability

Another issue is the avaliability. For almost all companies, having their services not avaliable to them, especially for a longer period of time, is a nightmare. It could cost them enormous amounts of money. One possibility are outages by the provider experiencing DDoS or other attacks. As said before, since the provider is a bigger target, attacks on the provider may also affect you. Another avaliability issue would be the financial situation of the provider. If the provider goes suddently bankrupt and shuts down it's services, you have to migrate to another provider. If it happens suddenly, you could lose important or sensitive data. Legal issues can affect avaliability as well. In 2009, the FBI raided datacenters in Texas and confiscates hundreds of servers. [22] This can lead to data loss and giving sensitive data to unauthorized people.

# Chapter 5

# Evaluation of existing systems

We have already looked at the details of IaaS and how it works technically. Now it's time to give you some insights how such cloud infrastructures are implemented in real life and how resources can be provided and used by the consumers. Therefore we want to illustrate some concrete representatives who are are operating as a successful IaaS provider in practice in the following chapter.

## 5.1 Amazon Elastic Compute Cloud (Amazon EC2)

### 5.1.1 A first glance on Amazon's world

Amazon EC2 is part of the big Cloud Platform of the famous and globally well-known internet company Amazon.com, Inc., namely Amazon Web Services (AWS). AWS started offering IT infrastructure services to businesses using this Cloud computing model already in 2006. Today it's probably the most popular representative in this kind of IT sector offering highly reliable, scalable and low-cost infrastructure platform in the cloud used by a huge number of businesses in currently 190 countries around the world. They emphasize and try to implement the several benefits Cloud computing brings with it. Consumers should avoid hight initial costs to get their infrastructure running, instead a variable cost model with a pay-as-you-go price model should make their work much more cost-effective. Furthermore setting up and maintaining infrastructure components


Figure 5.1: Amazon Web Services

should not be a common task businesses should care about. In contrast the focus should be set on the major issues regarding the concrete business and infrastructure capacity can be provided by AWS with just a click on-demand. AWS truly offers a broad variety of several different services distinguished by it's provided resource type. A typical key benefit someone often face when talking about and using AWS is the perfect interoperability of all the platform components comprising it. That's probably a major point what makes AWS so beloved by customers around the globe. Next to storage(Amazon S3) or networking (Amazon VPC) compute resources, managed by Amazon EC2 are essential to every business as these are the physical machines services need to be alive [14].

### 5.1.2 EC2 Insights

Amazon EC2 is designed for web and system administrators to make their lives easier. It provides easy manageable compute capacity available in the cloud. Via a web interface the administrator can request more or less capacity with just a few clicks within minutes. Some more minutes later the required instances are fully set-up with the desires of the user and ready-to-go. This allows the user to scale up and down (which means renting more or compute capacity) on the fly depending on the current demand and without any need to configure them manually. To give this explanation further importance and motivate why Amazon EC2 is the right choice for so many businesses, we want to look at some key benefits this cloud technology introduces in contrast to buying just traditional dedicated hardware at the provider of your trust.

#### 5.1.2.1 Elastic scaling

As already mentioned the customer can add and remove EC2 instances as she likes and therefore adapt the available compute capacity based on the current demand. It doesn't matter whether someone commissions just a few or even thousands of instances simultaneously [14] Especially for companies whose resource demand is highly unpredictable due to the reason that they can't give a good guess about the future development of their product this EC2 technology can help them a lot. A question which probably arises is what happens if workload on the EC2 instances varies heavily due to the nature of the application. For instance there could be the case that load heavily depends on some special days or time and always manually enlarging and shrinking the cloud infrastructure would be again an overhead. For this reasons Amazon introduced the Auto Scaling feature where the capacity dynamically scales up and down depending on the current resource demand. The user has opportunities to define peaks and troughs or specific time points where EC2 should react with scaling. So on the one hand you can ensure that there is always enough capacity to deal with the current load and on the other hand you can minimize your costs as you don't waste money for capacity you don't even need at the moment. Over the years Amazon developed some further possibilities for customers to get the most "bang for the buck". With the so called Reserved Instances they established a model where customers can agree to select one of the predefined instance types, pay some low initial costs for each instance they want to reserve and get a significant discount in variable costs. If you are not satisfied with your instances anymore you still have the possibility to place them in another AWS region, change the instance type or even sell capacity to other projects that end before the time frame for the Reserved instances expires. A further pretty nice instance model are Amazon EC2 Spot instances. With Spot instances the user can bid on the hourly price per instance she want to spend. Compute capacity again increases and decreases dynamically and as long as the bid meets or exceeds the instance price which customers commit to pay they gain access to them. This is a great possibility to be sure not to pay more than you actually want and prevent the risk that you oversee massive scaling [14].

#### 5.1.2.2 Elastic Load Balancing (ELB)

ELB is an intelligent solution to improve both scalability and fault tolerance. The load balancer component automatically distributes incoming traffic throughout the available instances to balance the load equally. It seamlessly integrates with the Auto Scaling feature described in the section above. You can still set your Auto Scaling conditions and if they are met instances get added to your Auto Scaling Group. This also works fine behind an Elastic Load Balancer. Achieving better fault tolerance is the second core point of using ELB. The Amazon load balancer automatically detects unhealthy instances behind itself and therefore does not distribute any further traffic to them until they get healthy again. So it's basically no problem when one of several instances is down for a while. Amazon lets you spread your EC2 instances within Availability Zones and multiple regions. A region is a specific geographic zone where Amazon operates and contains multiple isolated Availability Zones. Especially for globally operating bigger businesses it is highly recommended to distribute instances throughout several Availability zones and multiple regions as this results in lower latencies and further increases fault tolerance. Although a complete availability zone is offline traffic can be routed to farther away regions intermediately. The nice integration with other Amazon Service such as the new Amazon Route 53 even enables DNS failovers in case the load balancer itself suffers [14].

#### 5.1.2.3 AWS Management Console

Moreover what makes working with AWS so convenient is how less effort you need to complete such tasks. The AWS Management Console does not require much expertise in system or server administration as it comes with a comfortable graphical user interface along. There you can choose which of the dozens of services you would like to look at and also have root access to any of your EC2 instances and so you can start, stop, check and manage them as you like.

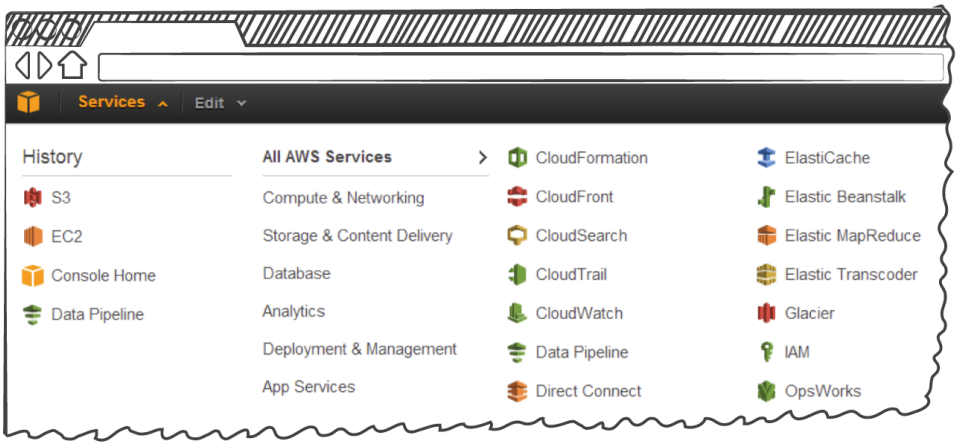


Figure 5.2: AWS Management Console - Main overview

## 5.2 Google Compute Engine (GCE)

### 5.2.1 Overview

Although the IaaS technology has been present and well-used for more than ten years and although Google is maybe the most successful internet and networking company, they stayed out of the IaaS market until 2013. Before that date Google of course was productive in the cloud world as well but focused more on the higher level of the cloud stack. With the Google App Engine they started to provide a convenient PaaS model which has become very successful due to it's simplicity and effective use within their cloud environment called Google Cloud Platform. Especially developers using Google's developing technologies love it because it nearly seamlessly integrates with them and allows to set-up a working platform in minutes. Newer trends showed that cloud services are also accepted as a potential and encouraging paradigm to solve scientific problems in addition to fulfil economic and business requirements. Cloud providers start to pay more attention regarding this further issue to make their current cloud model more scientific-oriented by adapting and changing their infrastructure. Drive by these new requirements Google came up with it's Compute Engine in 2013 as a new competitor on the market for established ones such as Amazon or IBM [23]. They officially describe it as designed to run high-performance, computationally intensive virtual compute clusters. As Google is a well-established big player in the networking world their global infrastructure make it a good candidate. The big palette of Google Services such as Google Cloud SQL and Google Cloud Storage moreover allow a nice integration [16].


Figure 5.3: Google Cloud Platform

As Google wants to be a strong competitor in a dense global market the main things offered are pretty similar to Amazon's EC2 for instance. In addition to on-demand scaling, load balancing or fault-tolerance Google emphasizes the use of local SSDs which evidently increases input-output operations per second and very low latency compared to traditional persistent disks. GCE offers a billing model broken down on a paying per-minute model after a 10-minute minimum charge always rounded up to the next minute and therefore promise the customer more flexibility. As a drawback of GCE the administrative opportunities are mostly mentioned as main limitation [9].

### 5.2.2 GCE for Scientific Computing

In the following section the focus is on the field of scientific computing where GCE claims to be very efficient. Investigating performance regarding Compute Engine surprisingly reveals some disadvantages compared to EC2 in terms of communication performance between data centres. Analyzing memory in contrast shows a different picture as GCE displays stronger memory performance against EC2. Especially storage transaction speed of each GCE instance identified as considerably faster as an EC2 instance. The storage data throughput again illustrates the speed in writing data whereby the general reading throughput is similar to Amazons infrastructure [23]. As a result of these evaluations we can see that GCE is well-suited for big data, data warehousing, high performance computing and further analytic applications [16].

# Chapter 6

# Conclusions and Future Work

In this article we have given an overview on the cloud type Infrastructure-as-a-Service (IaaS) allowing on-demand provision of compute, storage and networking resources via a software virtualized environment. When deploying applications on the global network, leveraging cloud computing facilities are a key factor for many successful world-wide operating business and IaaS provides the fundament where platforms and software products can be built on. A huge variety of features, simple configurability and several big internet companies as providers make IaaS a solid solution and attractive for any type of internet business. The flexibility and elasticity of compute resources with their on-demand and pay-what-you-need strategy make it easy to adapt to the current situation and workload and removes some annoying limitations of traditional physical dedicated hardware solutions. Due to software-based management and simple web interfaces as popular IaaS providers like Amazon EC2 or Google Compute Engine (GCE) provide it, the times of setting-up a dedicated server using vim or nano to manually change some SSH options are history. By just a few clicks new infrastructure instances or data centres are ready and running letting companies focus on their business related tasks and nothing else.

IaaS has moved away from a hyped trend to a well established cloud technology and is sill under ambitious development. Popular providers like AWS or GCE are running big projects and try to further improve simplicity of configuration tasks, high performance infrastructure clusters and availability.

# Bibliography

[1] Emmanuel Arzuaga and David R. Kaeli. Quantifying load imbalance on virtualized enterprise servers. In *Proceedings of the First Joint WOSP/SIPEW International Conference on Performance Engineering*, WOSP/SIPEW '10, pages 235–242, New York, NY, USA, 2010. ACM.

[2] A. Beloglazov and R. Buyya. Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints. *Parallel and Distributed Systems, IEEE Transactions on*, 24(7):1366–1379, July 2013.

[3] Liuhua Chen, Haiying Shen, and Karan Sapra. Distributed autonomous virtual resource management in datacenters using finite-markov decision process. In *Proceedings of the ACM Symposium on Cloud Computing*, SOCC '14, pages 24:1–24:13, New York, NY, USA, 2014. ACM.

[4] SSH Communications Security Corp. User authentication with certificates. `https://support.ssh.com/manuals/server-admin/44/User_Authentication_with_Certificates.html`.

[5] Dialogic Corporation. Introduction to cloud computing, jul 2010.

[6] Oracle Corporation. Making infrastructure-as-a-service in the enterprise a reality, apr 2012.

[7] Rodrigo de Souza Couto, Stefano Secci, Miguel Elias Mitre Campista, and Luis Henrique Maciel Kosmalski Costa. Network design requirements for disaster resilience in iaas clouds. In *Communications Magazine IEEE*, pages 52 – 58, oct 2014.

[8] A. Gajbhiye and K.M.P. Shrivastva. Cloud computing: Need, enabling technology, architecture, advantages and challenges. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference* -, pages 1–7, Sept 2014.

[9] Google. Compute engine scalable, high-performance virtual machines. `https://cloud.google.com/compute/`. Accessed: 2015-12-13.

[10] Google. Security in google's datacenters. `https://www.google.com/about/datacenters/inside/data-security/index.html`.

[11] He Huang and Liqiang Wang. P amp;p: A combined push-pull model for resource monitoring in cloud computing environment. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pages 260–267, July 2010.

[12] Apprenda Inc. Iaas, paas, saas (explained and compared). `https://apprenda.com/library/paas/iaas-paas-saas-explained-compared`.

[13] Kabooza. Kabooza global backup survey. `http://www.kabooza.com/globalsurvey.html`, jan 2009.

[14] Sajee Mathew. Overview of amazon web services, nov 2014.

[15] Wayne A. Jansen; NIST. Cloud hooks: Security and privacy issues in cloud computing, 2011.

[16] N. Serrano, G. Gallardo, and J. Hernantes. Infrastructure as a service and cloud technologies. In *Software, IEEE*, pages 30 – 36, apr 2015.

[17] IBM Global Technology Services. Resilience in the era of enterprise cloud computing, apr 2013.

[18] Michael Sonntag. Cloud security - an introduction, 2014.

[19] Stefan Bachmair; Christoph Stenglein. Comparison of cloud account security: Automatic management, dec 2014.

[20] Trendmicro. World backup day: The 3-2-1 rule. `http://blog.trendmicro.com/trendlabs-security-intelligence/world-backup-day-the-3-2-1-rule/`, apr 2013.

[21] Jonathan Stuart Ward and Adam Barker. Self managing monitoring for highly elastic large scale cloud deployments. In *Proceedings of the Sixth International Workshop on Data Intensive Distributed Computing*, DIDC '14, pages 3–10, New York, NY, USA, 2014. ACM.

[22] Kim Zetter. Fbi defends disruptive raids on texas data centers. `http://www.wired.com/2009/04/data-centers-ra`, sep 2009.

[23] L. Zheng Li, OBrien, r. Ranjan, and M. Zhang. Early observations on performance of google compute engine for scientific computing. In *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, pages 1 – 8, 2013.