Yekaterina Ladeichshikova

Implementing a cloud gaming solution with Amazon Web Services

Bachelor's thesis
Information Technology

2019

South-Eastern Finland
University of Applied Sciences

South-Eastern Finland University of Applied Sciences

Author (authors)	Degree	Time
Yekaterina Ladeichshikova	Bachelor of Engineering	May 2019

Thesis title

Implementing a cloud gaming solution with Amazon Web Services

40 pages

Commissioned by

Supervisor

Matti Juutilainen

Abstract

The purpose of this study was to learn more about cloud computing and cloud gaming, to implement a personal cloud gaming service and to analyse the results via gameplay testing. The process of the study included investigating the advantages of cloud gaming, comparing different cloud computing services that could be used and attempting to create an efficient cloud gaming system.
The practical part of the study included assembling a virtual server for the service, choosing the suitable client-side application and setting up the connection between them. The tools that were used were Amazon Web Services as a cloud service provider, a gaming platform with the ability to stream games, remote desktop application and a VPN (Virtual Private Network) software.

Keywords

Cloud gaming, cloud computing, Amazon Web Services, cloud server

CONTENTS

1 INTRODUCTION

The growth of video games industry results in a requirement of more powerful hardware. Gamers meet a problem of constantly upgrading their machines in order to play modern games. They have to spend immense amounts of money not only on games, but also on necessary computer parts. Cloud gaming, the technology based on cloud computing infrastructure that runs a game on the server and streams it to game consoles or personal computers, can be a perfect solution to this issue (Techopedia 2019). Cloud gaming also works in the interest of game developers who don't have to port their games to different platforms.

The process behind cloud gaming is collecting users' inputs and actions, transferring them to the cloud server, processing them, rendering the result image, compressing it and streaming it back to the users' machines. All these actions are executed for every event. Therefore, keeping the latency low enough for a good gaming experience can be very challenging.

The purpose of this study is to learn more about cloud computing and cloud gaming, to implement a personal cloud gaming service and to analyse the results via gameplay testing. The process of the study includes investigating the advantages of cloud gaming, comparing different cloud computing services that can be used and attempting to create an efficient cloud gaming system. The ready solution needs to have low enough latency and satisfactory bandwidth to play various games.

The practical part of the study includes assembling a virtual server for the service, choosing the suitable client-side application and setting up the connection between them. The tools that are used are Amazon Web Services as a cloud service provider, a gaming platform with the ability to stream games, remote desktop application and a VPN (Virtual Private Network) software.

2 CLOUD COMPUTING

In order to understand cloud gaming, it is important to know the definition of cloud computing. Cloud computing can be defined as a service that provides computer resources and storage space that are accessed via the Internet (Murugesan & Bojanova 2016, 3). Cloud provides an end user with scalable infrastructure for application hosting or storing data. Cloud computing became an efficient solution by reducing the cost and physical storage space. The main goal of cloud computing is accessibility, meaning that it provides access to any computational requirements to any location at any time.

2.1 Cloud computing methods

This section is based on the cloud computing types understanding of Murugesan and Bojanova (2016, 6–7). Each cloud service provider has a different set of features it offers depending on the customer's needs. Cloud computing services can be divided into three types as shown in Figure 1. These types give different amount of control over the cloud and have specific functions.

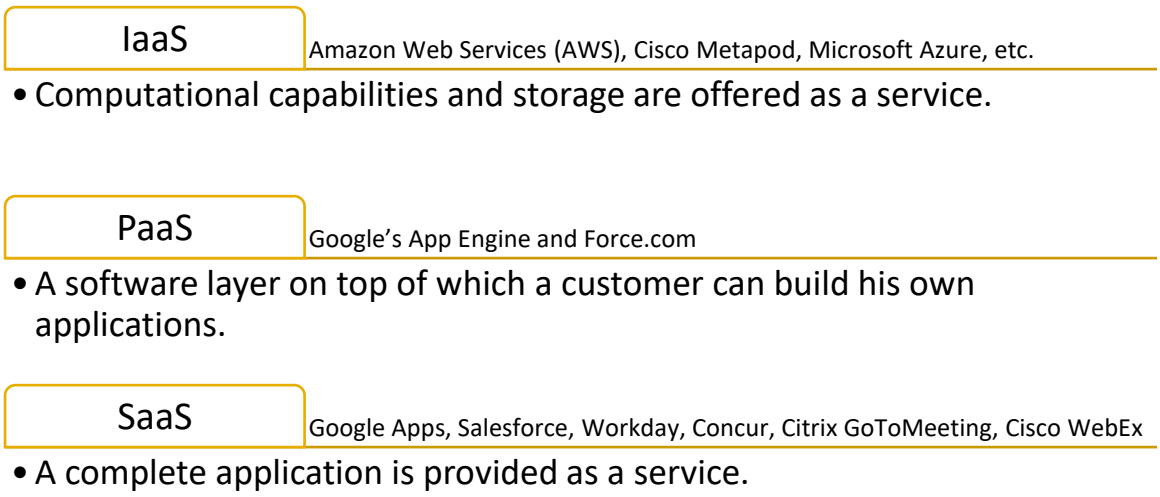

Figure 1. Cloud models (Torry Harris 2019)

The following list explains the purpose of each type:

1. **Infrastructure as a Service (IaaS)** is a type of cloud computing service where a provider offers an end user an infrastructure. Computer

infrastructure is a collection of computational capabilities and storage. Users have a freedom of deploying any necessary software and using resources in a way needed. Examples of IaaS providers are Amazon Web Services (AWS), Cisco Metapod, Microsoft Azure, Google Compute Engine (GCE) and Joyent.

2. **Platform as a Service (PaaS)** is a software layer on top of which a customer can build his own applications. In other words, it is a development environment offered as a service over the Internet. Usually PaaS provides users with an operating system and some predefined software for application development. Two of the known PaaS examples are Google's App Engine and Force.com.
3. **Software as a Service (SaaS)** offers a customer a complete application as a service. The application is hosted in the cloud and can be accessed simultaneously by multiple users. It is beneficial for both parties as customers don't need to manage the servers and providers reduce their expenses by running only one instance of the application. Google Apps, Salesforce, Workday, Concur, Citrix GoToMeeting, Cisco WebEx are some of the SaaS examples.

In this study the emphasis is on the IaaS model as it is most suitable for the given goals.

2.2 Benefits

Cloud computing became an optimal solution for various organizations, including small businesses and bigger companies. It has several advantages compared to the classical server approach. The following list shows at least some benefits of cloud computing based on the article by Bozicevic (2018):

1. **Cost savings**. Cloud computing allows small companies save their money on actual server equipment and instead delegate all the work to the cloud service. Cloud subscriptions tend to appear cheaper than building a physical server environment. Furthermore, it prevents the company from having the unnecessary workload. It allows them to focus on the requirements they need and choose the corresponding cloud service.

2. **Reliability**. Cloud services provide a certain level of security. It is their responsibility to keep the stored data safe and inaccessible to unauthorized individuals. Moreover, it is made sure that the data won't be lost as cloud service providers offer loss prevention. However, in case of an emergency situation, cloud-based services provide a quick data recovery.
3. **Flexibility**. Keeping the data or workflow in the cloud doesn't take any physical space in the office. Companies are able to extend their business without worrying about buying additional IT equipment. Flexibility of cloud computing improves the efficiency and general user experience of the company.
4. **Accessibility**. A cloud service can be accessed at any time from any location. This can be a significant factor for many types of organizations. Additionally, it is a convenient option for remote employees who may work from different time zones.

These are only some of the benefits that suggest that adopting cloud computing approach is more advantageous.

2.3 Challenges

Although cloud computing proved itself to be an effective and beneficial service, it often meets some hesitation coming from various enterprises. Therefore, there are several challenges for cloud service providers to guarantee the faultless user experience for their customers.

Data security is an important aspect in the cloud service provider's responsibilities. Organizations rely on the provider to keep their invaluable information safe. The cloud must choose a reliable encryption method, make certain that their hardware is properly secure, arrange disaster recovery in case of accidental data loss.

The other thing the cloud service provider should keep in mind is a billing method. Usually cloud users are charged on a pay per usage basis, but the billing

method can vary. It may depend on a cloud computing type, whether it is IaaS, PaaS or SaaS. Cloud service providers should adopt a suitable billing solution that meets their needs and find a trustworthy billing management software.

According to Hwang (2012, 225–227) there are six main architectural design challenges in cloud computing development. They can be listed as follows:

1. **Service availability**. Achieving high availability can become an issue in the cloud server environment. Having several datacentres in different locations doesn't always guarantee the flawless workflow as they tend to have common software infrastructure. Using multiple cloud providers is a possible solution to avoid failure.
2. **Data privacy and security concerns**. As it was mentioned earlier, data safety is highly important in cloud computing. The fact that most of the modern cloud providers offer public clouds leads to more potential threats. Some of the attacks can be easily prevented by adopting common security technologies as encryption, virtual LANs, firewalls etc. One of the options is to encrypt the data before placing it in a cloud.
   Data security, however, is not the only concern. Clouds can also be targeted by traditional network attacks like DoS attacks, spyware, malware, rootkits, Trojan horses and worms. Passive attacks are used to steal passwords and sensitive data. Active attacks are focused on causing damage to cloud servers.
3. **Unpredictable performance and bottlenecks**. I/O sharing in cloud computing can be problematic, unlike CPU or memory sharing. The process of writing on a physical disk significantly reduces the latency. In order to improve I/O architecture it is necessary to efficiently virtualize interruptions and I/O channels.
   Bottlenecks is another cloud computing issue. As Internet applications are currently very data-intensive, data transport and placement across the cloud boundaries can be challenging. Therefore, cloud providers should eliminate all the bottlenecks, widen bottleneck links and remove weak servers.

4. **Distributed storage**. In cloud computing data centres should meet the requirements of scalability, data durability and high availability. The reason for it is that cloud databases are always expanding, and they demand corresponding scalability from the storage system. It should be able to meet the cloud's possibility to scale both up and down on demand. The solution for that is efficiently distributed Storage Area Network (SAN).
5. **Cloud scalability and interoperability**. Cloud providers' billing method is usually pay-as-you-go model which applies to storage and network bandwidth. Therefore, cloud should be able to scale quickly according to load variation.

   In order to achieve interoperability in the cloud computing environment it is necessary to have a possibility to distribute virtual machines between platforms. Distributing VMs and software can cause an issue of incompatibility. Open Virtualization Format (OVF) is an open, secure and efficient way of VM packaging and distribution. It does not rely on a specific host platform, virtualization platform or guest operating system.
6. **Software licencing**. The licencing model of commercial software is not ideal for cloud computing purposes. Therefore, cloud providers sometimes prefer open source software. Unfortunately, open source software is not very popular. The solution is improving licencing structure of commercial software companies to better fit the cloud computing model. (Hwang 2012.)

All these challenges should be considered when designing and managing a cloud computing service. It is highly important to keep them in mind and make sure that there are no vulnerabilities in the cloud.

3 CLOUD GAMING

Cloud gaming is a new branch of cloud computing that allows people to play games from remote cloud servers (Techopedia 2019). The purpose of cloud gaming is allowing anyone to play games on any type of computer. The growth of gaming industry results in games of very high quality which demand more

powerful hardware. Cloud gaming eliminates the problem of constantly upgrading the computer's hardware, as it offers a gaming platform as a service. Cloud gaming service has servers built specifically for gaming purposes, and the only thing the customer needs is a proper Internet connection.

Cloud gaming has many advantages not only from customers' perspective. Game developers also benefit from cloud computing, as it removes the necessity of porting games to different platforms. This process consumes time and resources and cloud computing is an efficient solution to the issue. Cloud gaming also eases the matter of incompatible hardware or software.

3.1 Background

Although gaming community still hasn't fully accepted cloud gaming concept, it has been around for almost two decades. It has been drastically improving since the first release of cloud gaming service, but continues to meet challenges and criticism coming from gamers.

Figure 2. OnLive cloud gaming service

The first cloud gaming demonstration was in 2000 at E3 conference by G-cluster company (D'Argenio 2018). It wasn't cloud gaming, as it's known nowadays, but it

was a start. It was streaming games over Wi-Fi to handheld devices. The next official launch of a cloud gaming system was by OnLive company (Figure 2). Steve Perlman, the CEO of OnLive, announced the service in March 2010 in his blog and presented it at E3 in June 2010 (Perlman 2010). Nevertheless, OnLive also had problems with latency and wasn't met with excitement, but with doubt.

In February 2011 another company, Gaikai, announced their version of the cloud gaming service. They offered to the customers to stream games like Mass Effect 2, Dead Space 2, Spore, Sims 3 and Second Life. According to Gaikai's CEO David Perry (2011), their service was able to run high-performance games at 60hz.

In 2014, Sony announced its own cloud gaming service PlayStation Now. It was based on the Gaikai technology after Sony purchased Gaikai for 380 million USD. Sony was offering various PlayStation games accessible from the cloud. The appealing feature of PS Now is the possibility to play PlayStation 2 and PlayStation 3 games on a PlayStation 4 console without porting them. Furthermore, it allows playing the games not only on PlayStation 4, but also on televisions, laptops, smartphones and other devices. First, it was launched only in the United Stated, but through the years it became available in Canada, the United Kingdom, Ireland, Japan and several European countries. In March 2019, PS Now service came to Finland.

3.2 Architecture

The idea of cloud gaming is that the game is streamed from the cloud server and it doesn't need to be installed on the user's computer. However, there is a complicated process behind it. Figure 3 shows the structure of cloud gaming.

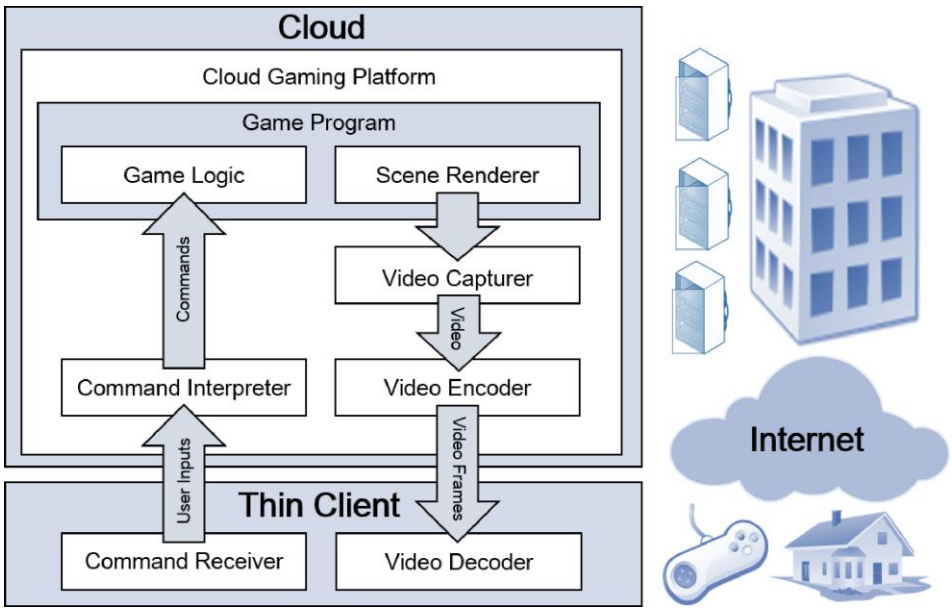

Figure 3. Cloud gaming framework (Cai et al. 2016)

As Figure 3 shows, client monitors user inputs and sends them through the Internet to the cloud. The cloud game platform receives these commands which are then interpreted into game actions that affect the game world. After the changes are made, they are processed by graphical processing unit (GPU) in order to render the scene. The render scene is encoded into video frames that are streamed back to the client where it's decoded and can be seen by the user. (Shea, Liu, Ngai & Cui 2013.)

3.3 Advantages

Information in this section is based on the studies of Huang et al. (2013) and Cai et al. (2016).

The interest towards cloud gaming is growing rapidly. As the history of cloud gaming shows, even big gaming corporations are starting to adopt cloud gaming. Gamers become more open to the idea of playing games from a remote source. There are several benefits of cloud gaming that will make gamers' lives easier:

1. **No upgrading of hardware**. Game developers make their games based on the latest and most powerful hardware in order to improve the quality. In the classical gaming approach gamers would need to keep up with the game market by purchasing most recent computer components. However,

as it was mentioned previously, in cloud gaming the game itself is running on a remote cloud server. This removes the necessity of constant hardware upgrades and prominently reduces the cost of playing video games.

2. **Playing games on different devices**. Cloud gaming allows playing games on various platforms and devices. Even the most demanding games can be played from laptops or even mobile phones. This is very appealing for those users who can't afford having powerful personal gaming computers. Moreover, it provides an opportunity to play favorite games while traveling.
3. **Enables playing more games**. As the user's choice of games is no longer limited by the hardware he owns, it provides a possibility to play more various games. Some cloud game providers offer a library of games included in the monthly subscription. Others allow playing any games that the customer owns.
4. **Migrating across different client devices during a game session**. One of the features of cloud gaming is the ability to switch between the devices and to continue the game on another one. For example, Stadia, Google's newest cloud gaming platform, allows continuing the game instantly on any device from the point where the user has stopped playing (Stadia 2019).

Additionally, cloud computing can be a solution for some of the issues that game developers are struggling with. Focusing on the cloud can work to the developers' advantage. The reasons for game developers to consider cloud gaming are following:

1. **Less software/hardware compatibility problems**. If developers create their game specifically for the cloud, the issue of software and hardware incompatibility disappears. There is no necessity of porting the games to different platforms, as all games are running on one platform on the cloud server.
2. **Reduced production cost**. For the reasons mentioned earlier, the production cost is reduced. Without the need of making separate versions of the game, developers are saving a considerable amount of money.

3. **Avoiding piracy**. Piracy is one of the biggest concerns of game developers. Cloud gaming prevents piracy, as all the games are safely stored on the cloud server and can't be illegally shared.

These benefits show the advantage of cloud gaming and the reason for users and developers to consider it as a beneficial gaming platform.

3.4 Challenges

Despite all the advantages cloud gaming offers, there are certain challenges that cloud gaming providers must overcome in order to provide the best Quality of Service (QoS). The complexity of cloud gaming architecture causes some issues that, without proper consideration, may degrade user experience. Users expect high-quality video and fast response rate from the gaming experience.

According to Shea et al. (2013) the two main challenges of cloud gaming are network latency and video encoding. High network latency may result in response delays that are not tolerable in some types of games. As shown in Table 1, various game types have different delay tolerance. Delay tolerance is the maximum response time a user can tolerate before it affects the Quality of Experience (QoE).

Response delays in first player shooters (FPS) may affect the game process and result in a player's disadvantage. FPS games are usually action-based, and the maximum acceptable delay threshold for them is 100 milliseconds. The response delay in third person games, such as role-playing games (RPG), is less crucial. In this type of games, the commands that the user enters are usually executed by the player's avatar, a figure representing the player in a video game. The delay before the action's execution may appear unnoticeable as the user doesn't expect the immediate response. However, if the game forces the player to wait more than 500 milliseconds and results in a negative outcome, it can be frustrating for the player. The third type of games is the most tolerant to response delays. Real time strategy games can have a maximum delay threshold of 1000 milliseconds without degrading the gaming experience. These games usually

execute several commands at the same time, and they are not expected to affect the game instantly. Longer response delay in RTS games may not be notices by the player.

Table 1. Delay tolerance in traditional gaming (Shea et al. 2013)

Example Game type	Perspective	Delay Threshold
First Person Shooter (FPS)	First Person	100 ms
Role Playing Game (RPG)	Third Person	500 ms
Real Time Strategy (RTS)	Omnipresent	1000 ms

The interaction delays mentioned apply only for the classical gaming approach. Cloud gaming makes the issue even more complex, as all the games are rendered from the cloud. It means that while in classical gaming response delay was important only in multiplayer games, in cloud gaming interaction delays occur even in single-player games.

Another cloud gaming concern is video encoding. In the process of cloud gaming the game video is streamed from the cloud server to the thin client on the user's computer. The video must be quickly encoded and delivered to the client. Encoding should be done with respect to a small number of frames, considering that the following frames are not obtained until the certain action is done on the client side. In order to avoid the response delay, the encoding process must be very efficient. The encoder must be chosen carefully in order to meet the requirements or cloud gaming's real-time encoding. (Shea et al. 2013.)

3.5 Future of cloud gaming

Cloud gaming went through a lot of changes and challenges. Despite all the doubt coming from different sources it keeps overcoming the obstacles and developing further. Cloud gaming has a potential future of affecting gaming industry and attracting more companies and customers. According to Cai et al. (2016), cloud gaming will go through dramatic changes and has several possible outcomes in the future.

There is a high probability for commercial cloud gaming services to start using context-aware optimization algorithms for cloud gaming videos. The idea of this algorithm is using in-game context, for example camera location and orientation, in order to improve interaction latency and graphic quality. This method provides an interface between the game engine and the video encoder, which means that the game itself doesn't need to be re-written. (Semsarzadeh et al. 2014.)

Another change that cloud computing may go through is developing new pri model. Currently there are three main charging models: subscription, per-game, and free to gamers (Cai et al. 2016). However, all of them have disadvantages and do not provide the best financial practice. The subscription model requires sufficient number of subscribers. The per-game model may not be convenient enough for users as the cost of renting a game is close to the price of the original game. The free model may appear to be risky considering that the customer can decide not to buy the game after the free trial. Therefore, commercial cloud computing services should possibly look into new pricing models that will be more beneficial for all the parties.

Multiplayer games may use the advantages of cloud gaming for the benefit of both services. The possible concern of the users about network connectivity in cloud gaming should not be a problem in multiplayer games, as those games originally require the network access. The ability of cloud gaming to run the game without installing it first can be highly appealing to multiplayer gamers. Furthermore, cloud gaming can provide competition fairness, meaning that players will have similar gaming experience. By removing the difference of purchased hardware and adapting the system to the terminal's network, it can provide more fairness.

Cloud gaming has a possibility to attract more observers and consequently more gamers, which can lead to new business models. The nature of cloud gaming allows streaming the game video to many observers easily without charging any fee. Gamers are interested in their matches being watched. The cloud infrastructure additionally enables interaction between gamers and observers.

These interactions, however, may appear to be problematic, if any of the observers are getting a delayed picture. This means that broadcast latencies should be taken into consideration.

The last forecast for cloud gaming is combining it with novel gaming paradigms such as virtual reality (VR) games and augmented reality (AR) games. Cloud computing has the possibility of boosting an interest in VR and AR games by offering rich rendering resources. Despite that, as it was mentioned in other cases previously, the network latency overhead is an open issue and should be considered in order to achieve a good gaming experience.

4 CLOUD SERVICES

There are many different cloud service providers on the market. Each of them has specific advantages and disadvantages. This chapter focuses on the two most popular services: Microsoft Azure and Amazon Web Services.

4.1 Amazon Web Services

Amazon Web Services (AWS) is a cloud computing service that provides various platforms to companies and individuals. At the moment, AWS is a known worldwide service that operates in 190 countries. It has several datacenters located in the US, Europe, Brazil, Singapore, Japan and Australia. According to Amazon website (2019), AWS has the following benefits:

1. **Low cost**. AWS works on pay-as-you-go pricing basis and offers affordable prices for their service. In the pay-as-you-go pricing model the user is billed for the resources that are used, not for the entire infrastructure (Techopedia 2019). This allows users or businesses to adapt depending on their needs and not overpay for unused computational or storage resources.
2. **Agility and instant elasticity**. AWS allows their users to quickly assemble their cloud servers without worrying about actual hardware. The work can be started instantly and then scaled as the workload grows. Any amount of

resources can be used for as long as needed. The customer will only pay for what is used.

3. **Open and flexible**. AWS is not limited to one platform or one programming language. Customers are free to use any service according to their business needs. That allows focusing on the workflow.
4. **Security**. AWS has industry-recognized certifications and audits: PCI DSS Level 1, ISO 27001, FISMA Moderate, FedRAMP, HIPAA, and SOC 1 and SOC 2 audit reports. That makes AWS a secure and reliable service. AWS datacenters are physically secured and offer high data integrity and safety.

For additional redundancy and reliability, the AWS infrastructure is built around Regions and Availability Zones (AZs). There are two types of Regions: AWS Regions and AWS Local Regions. Each AWS Region provides several isolated Availability Zones in different locations. These AZs are connected with low latency and high throughput, which provides a necessary redundancy and availability to the network. An AWS Local Region is designed for a disaster recovery option for a certain AWS Region. It is a single datacenter region and is isolated from other AWS Regions. AWS has 61 Availability Zones in 20 geographic regions.

4.1.1 Amazon Elastic Compute Cloud

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides scalable computational resources to the customer. It allows creating as many virtual servers as the user needs and having a full control over its environment, including security, networking and storage. The virtual environment in EC2 is called an instance. The user can own and manage multiple instances at a time. The instance is preconfigured with Amazon Machine Images (AMIs) that contain an operating system of choice and additional software. Every instance can be equipped with different hardware. The configurations of CPU, memory, storage and networking capacity are called instance types and are selected during the instance setup. EC2 instances have secure key-pair verification for the login. (AWS 2019.)

4.1.2 Amazon EC2 pricing

This section is based on the information from AWS website (2019). Amazon offers four different pricing models for the EC2 instances: On-Demand, Reserved Instances, Spot Instances, and Dedicated Hosts. For On-Demand instances the customer is charged for computing capacity by per hour or per second basis. Computing capacity can be scaled up and down depending on the needs of the application which would not affect the price. This type of instances is suitable for short-term application with unpredictable workloads. It can help avoiding long-term commitment and upfront payments.

Spot instances offer spare EC2 computing capacity for lower price than On-Demand instances. With spot instances the user is allowed to choose the amount he wants to pay for the instance. While the bidding price is lower than the actual instance price, the instance is running. Spot instances are a good solution for flexible applications that are not affected by interruptions.

Reserved Instances are cheaper than On-Demand instances, but they demand a commitment from the customer. They provide a capacity reservation, which is good for applications with steady state and predictable usage. The usage term of Reserved Instances is over 1 or 3 years.

A Dedicate Host provides the user with physical EC2 server. The cost is reduced by using the software licenses owned by the customer. It can either be purchased by the On-Demand pricing (hourly) or as a Reservation for lower price.
Amazon additionally offers a one-year free trial for Linux and Windows micro instances.

4.2 Microsoft Azure

Microsoft Azure is a set of cloud computing services created for organizations and businesses to help them build, test, deploy and test various applications (Microsoft Azure 2019). Microsoft Azure is a worldwide service. Azure is available in 140 countries and operates in over 54 Azure regions (Figure 4).

Figure 4. Azure global infrastructure (Microsoft Azure 2019)

Microsoft Azure (2019) website lists the following benefits of their services:

1. **Productive**. Azure provides more than a hundred different services and tools for the customer. It supports various programming languages and frameworks, including Node.js, Java and .NET.
2. **Hybrid**. Azure allows building the applications across both on-premises and cloud environments. It has low latency network by using fast hybrid connectivity through Azure ExpressRoute at 100 Gbps. It is secured with Azure DDoS Protection and Azure Firewall.
3. **Intelligent**. Azure offers pre-built APIs for building the artificial intelligence (AI) solutions, such as machine learning models. The services that Azure provides are Azure Databricks, Azure Cosmos DB, Azure Cognitive Services, and Azure Bot Service.
4. **Trusted**. Azure offers trusted and secure infrastructure. It is validated with Microsoft Cloud certifications.

Azure is a recognized service that can be used in different business fields as healthcare, financial services, government, retail, and manufacturing.

4.2.1 Azure Virtual Machines pricing

Microsoft Azure pricing models are similar to the ones AWS has. Azure operates using Pay-as-you-go and Reserved VM Instance models (Microsoft Azure 2019).

In Pay-as-you-go model the client is paying for the computational capacity by the second of usage. It doesn't demand any long-term commitment or upfront payment. The user can expand the capacity at any time and pay only for the resources used. It is useful for the applications that need flexibility and have unpredictable workloads. Users that are starting to use the service for the first time can also benefit from this pricing model.

An Azure Reserved Virtual Machine Instance is a possibility to purchase a VM for one or three years in a specified region. This pricing model demands a commitment from the client and an upfront payment. In return the client receives a significant discount of about 72% compared to the Pay-as-you-go pricing. This type of instances can be exchanged or returned, if needed. They are recommended for the steady applications with planned workload. Customers get budget predictability and are able to use the instance for a long time period.

5 IMPLEMENTING A CLOUD GAMING SOLUTION

This chapter describes the step by step process of implementing a cloud gaming server using Amazon Web Services. It shows the requirements that are needed to create a working solution. It thoroughly follows the setup process of a cloud server. In the end of the chapter the service is tested, and results are observed.

5.1 Requirements

The goal of this project is building a virtual gaming server, configuring a client machine and setting up a connection between them. For these purposes there are certain requirements that need to be met. They are the following:

- AWS EC2 instance

- Client computer
- Remote Desktop software
- Steam (or other service that supports game streaming)
- Virtual Private Network (VPN)

The AWS instance needs to meet specific technical requirements in order to stream high definition games. The graphics card is one of the most important hardware components for this purpose. AWS offers several GPU (Graphic Processing Unit) instance types that provide the instances with powerful graphics cards. G2 instances are good for graphic-intensive applications and appear to be suitable for the provided work. They are backed by NVIDIA GRID GPU (Kepler GK104) graphics card. For this project a g2.2xlarge instance type was chosen.

The specifications of g2.2xlarge instances as they are listed on Amazon website (2019) are the following:

- 26 ECUs (EC2 Compute Units)
- 8 vCPUs (2.6 GHz, Intel Xeon E5-2670)
- 15 GB memory
- 1 x 60 GB Storage Capacity

Asus ZenBook UX30FA is used as a client machine. Asus ZenBook is a laptop with an integrated video card that allows testing the features of cloud gaming. Originally, this machine is not able to run games with high graphics quality. In this study the ability is to play various types of games on a simple laptop tested. The specifications of the laptop are the following:

- Intel Core M 5Y10 Processor
- 4 GB memory
- Integrated Intel HD Graphics 5300
- 256GB SSD

Remote Desktop software is needed to connect to the virtual server in order to configure it. It is not used during the testing of the gaming process. The native

Windows application Remote Desktop Connection is used for this project. It is preinstalled in the Windows 10 operating system and easy in use.

Steam is a digital distribution platform created by Valve Corporation that also offers a social networking service. One of the features Steam provides is In-Home Streaming. It allows users to stream the games from one PC to another in a private network. This feature helps to create a cloud gaming experience for the project.

VPN is needed for the purpose of the study considering that both the client PC and the cloud server have to be in the same private network. VPN can help solve this problem, as it extends a private network over the public network. For this project ZeroTier One technology was chosen. ZeroTier One is an open source software that establishes Peer to Peer VPN connection between devices. It is easy to configure and use.

5.2 Instance setup

The first step to create a cloud gaming server is initiating an AWS instance. It is necessary to choose between Spot and On-Demand Instances. For this study the Spot Instance pricing model was chosen. The main reason for this is that the Spot Instances have lower price per hour of usage.

Spot Instances allow the customer to use the spare computational resources that AWS provides. It works so that the user submits a Spot Request. In the request the user specifies the parameters for the desired instance and the maximum price he agrees to pay for the instance. Spot Request is succeeded when the bid that the user made is higher than the actual Spot Instance price. While Spot Request is in a success state, the instance can be used. When the request is declined, the instance is either stopped or terminated; the user can choose the interruption behavior. This performance may appear to be a disadvantage for a cloud gaming server, as it can be interrupted at any point by AWS and can’t be stopped by the user. However, for the educational purposes of this study, this option remains the best. AWS requires the user to submit a special request for

On-Demand GPU instances, whereas with Spot Instances the user is free to use any instance type of his choice.

In order to start a Spot Instance, it is necessary to create a Spot Request. The user needs to specify the instance configurations. The first step is to choose the AMI (Amazon Machine Image). AMIs allow the user to start the work instantly without installing the operating system or worrying about basic configurations. For this project I have selected Microsoft Windows Server 2019 AMI. Windows is the most suitable option for a gaming server.

The next step is to select Minimum compute unit either as instance specifications or as an instance type. I chose the second option by selecting g2.2xlarge as the instance type. Then the user needs to specify the network to use and an Availability Zone. I used the default network that was preconfigured by AWS. This project doesn't demand a complex network structure; therefore, the default network applies well. In case of Availability Zone, it is better to select "No preference" option. This way the instance is balanced across all Availability Zones.

The next step is to create a key pair for the security. AWS requires to give the pair a name and then sends a private key file to the user. This file is used either for verification on the login or, in case of Windows instances, to obtain an administrator password to the OS. A single key pair can be used with multiple instances. I created a key pair with the name "game_server".

The user can add additional storage units to the instance or change the size of the existing one. I added a General Purpose SSD with a size of 60 GB. This SSD is needed to store the installed games. The system drive is also a General Purpose SSD with a size of 30 GB.

After that it is important to create a Security Group (Figure 5). Security Group is a set of inbound and outbound rules for the network. It allows blocking the undesired traffic to secure the network. It is important, however, to allow the

traffic between the client machine and the network. I created a Security Group and allowed TCP (Transmission Control Protocol), UDP (User Datagram Protocol), ICMP (Internet Control Message Protocol) and RDP (Remote Desktop Protocol) traffic flow from the client machine's IP address.

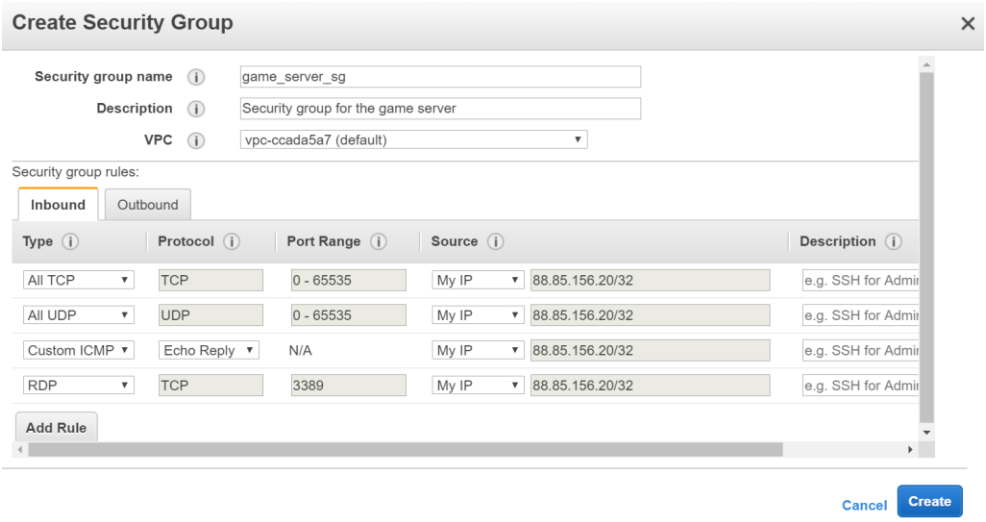

Figure 5. Security Group configuration

The next steps focus on configuring the Spot Request itself. First of all, the total target capacity and interruption behavior are specified. For this project I work with a single instance and I want the instance to stop when Spot Request is declined. Then the Fleet request is configured. The user can specify multiple instance types for the fleet and the target capacity is chosen from one of them. Having more types improves the chance of having the target capacity reached. However, in order to plan the final cost and specifications of the Spot Instance, I selected only g2.2xlarge instance type for the fleet.

The final step is to enter the maximum price for the Spot Instance. There is a default setting that sets the price to be equal to the On-Demand price of the instance. Another option is to specify a price of choice. I set the maximum price to USD 0.4 per hour which is a little higher than the original Spot price as can be seen in Figure 6.

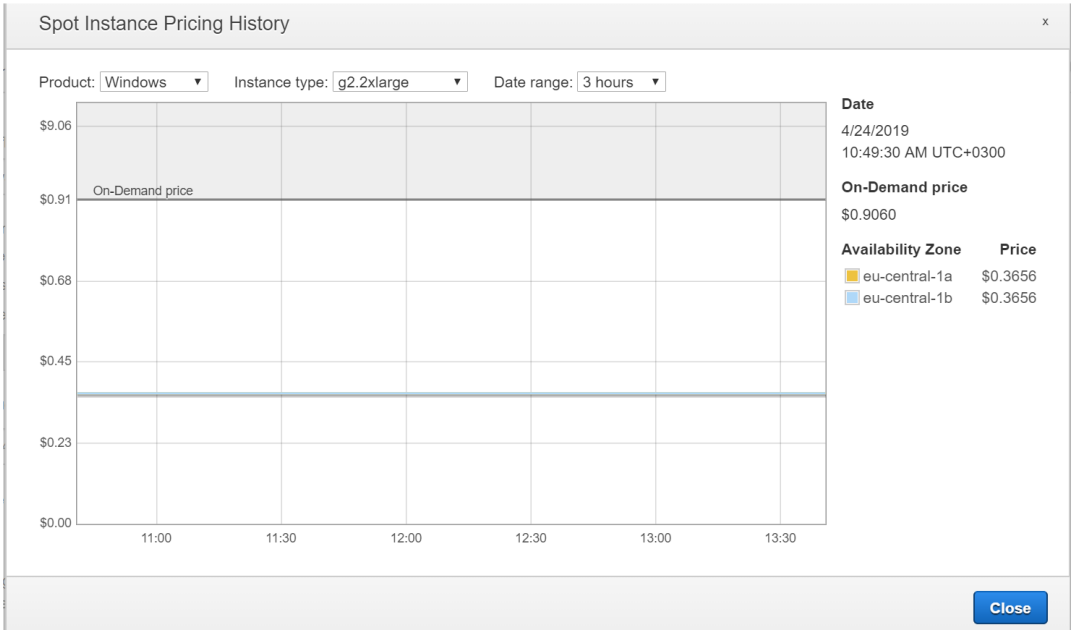

Figure 6. Spot Instance Pricing History

After that the Spot Request is ready and can be submitted. If the request is successful after the launch, the instance is created. Then the instance can be accessed with a Remote Desktop software by downloading an RDP file from the AWS dashboard. In order to login into the system the user needs to retrieve a password by submitting his private key and decrypting the password.

5.3 Software installation

After the instance is configured and started, the necessary software should be installed. Firstly, the video card drivers are downloaded. The drivers needed can be found on the NVIDIA (2019) official website. As can be seen in Figure 7, NVIDIA offers two options for finding a right driver. The first option suggests finding the driver manually by entering the product type, series and OS. The second option allows finding it automatically by scanning the system. This option is only supported by Internet Explorer and demands the latest Java version to be installed. I selected the second option and downloaded the driver for NVIDIA GRID K520 video card (version 360.35). After the driver is installed the system should be rebooted.

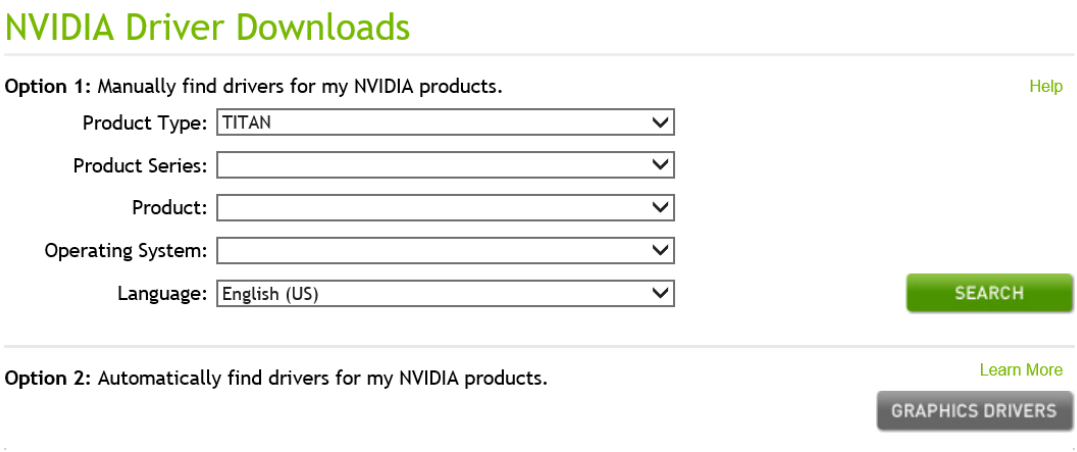

Figure 7. NVIDIA driver downloads page (NVIDIA 2019)

For the streaming, Steam is using H.265 video encoding. In order for it to work, NvFBC should be enabled. NVIDIA provides a NvFBCEnable tool. To enable it, I first disabled the Microsoft Basic Display Adapter and then ran the following command: *NVFBCEnable.exe -enable -noreset* (Figure 8). The system needs to be rebooted after that.

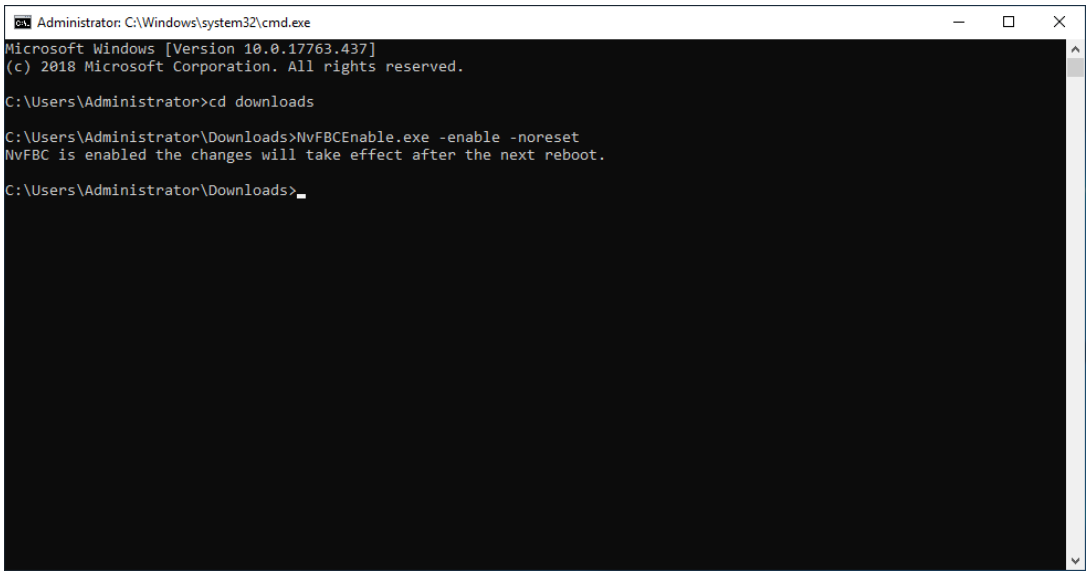

Figure 8. NVFBCEnable command execution

To ensure that games are using the video card, the Microsoft Basic Display Adapter should be completely removed from the system. In Device Manager it needs to be disabled and uninstalled (Figure 9). Then the following commands should be run in the Command Prompt:

- *takeown /f C:\Windows\System32\Drivers\BasicDisplay.sys*

- *cacls C:\Windows\System32\Drivers\BasicDisplay.sys /G Administrator:F*
- *del C:\Windows\System32\Drivers\BasicDisplay.sys*

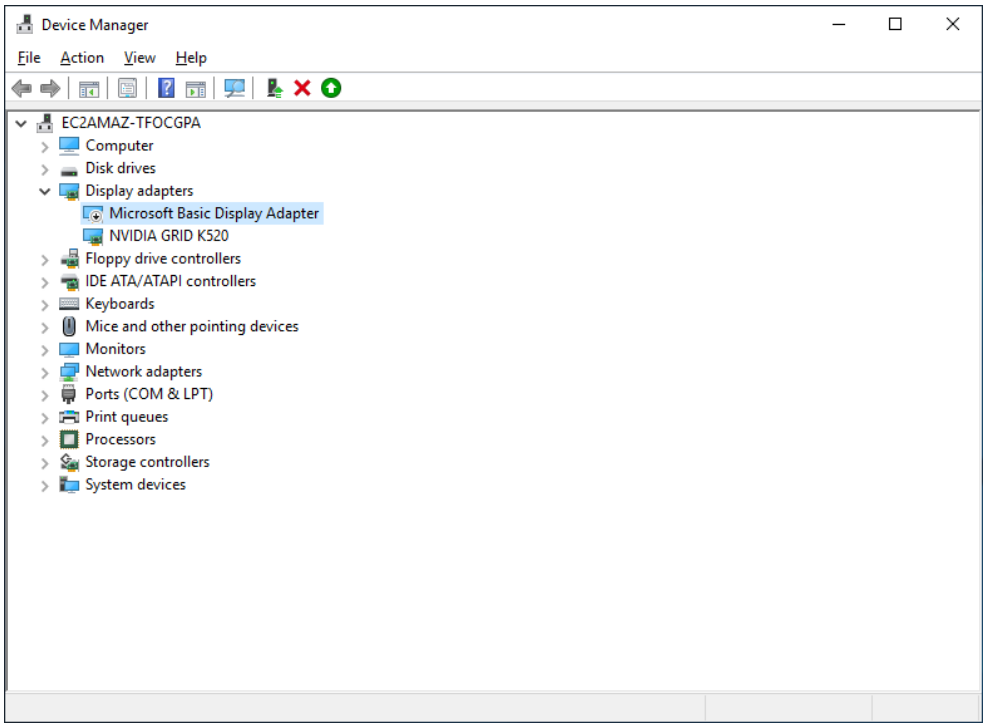

Figure 9. Disabling Microsoft Basic Display Adapter

The next step is to turn on the Windows Audio Service. It can be enabled in Windows Services. Additionally, as the EC2 doesn't have a soundcard, I needed to install a virtual one. I installed Razer Surround to get a sound simulation.

Afterwards, it is important to configure a VPN. For this project I chose the ZeroTier One software. As mentioned previously, ZeroTier One is an open source software that creates a Peer to Peer VPN connection between devices (ZeroTier 2019). It is free and easy to configure. First of all, it should be installed on both devices: the client machine and the cloud server. After the installation the service is started automatically. In order to connect the devices to the same virtual network it is necessary to create one. It can be done on the ZeroTier Central website. ZeroTier Central is a dashboard for configuring the network. Figure 10 shows the interface for creating a network. The user only needs to set the name for the new network and ZeroTier creates a unique ID. This ID allows devices to connect to the network. I created a network with the name

“game_network”. Then I created a pool of IP addresses to assign to connected computers.

Figure 10. ZeroTier network

The ZeroTier One application needs to be running on both machines and be connected to the created network. If connection is successful, machines get their virtual private IP addresses and are able to ping each other. The status and the IP addresses of all the connected devices can be checked from ZeroTier Central.

The last step is Steam installation. Steam can be downloaded from the official website. After the basic installation the following settings should be set:

- **Making Steam save the user credentials so that it can auto-login**. The server needs to be able to start Steam automatically without user intervention.
- **Setting the default folder for the downloads in Steam preferences**. In this project I had a separate drive for the games. I created a folder Steam Library in the root of the drive and set it as a default folder for Steam.
- **Enabling Steam In-Home Streaming**. In order for Steam to be able to stream games from the server, the streaming setting needs to be enabled. After enabling it, the device name should appear in the table as can be seen in Figure 11.
- **Enabling Hardware Encoding**. In order to improve the quality of streaming this option should be enabled. It can be found in *In-Home Streaming > Advanced Host Options > Enable Hardware Encoding*.

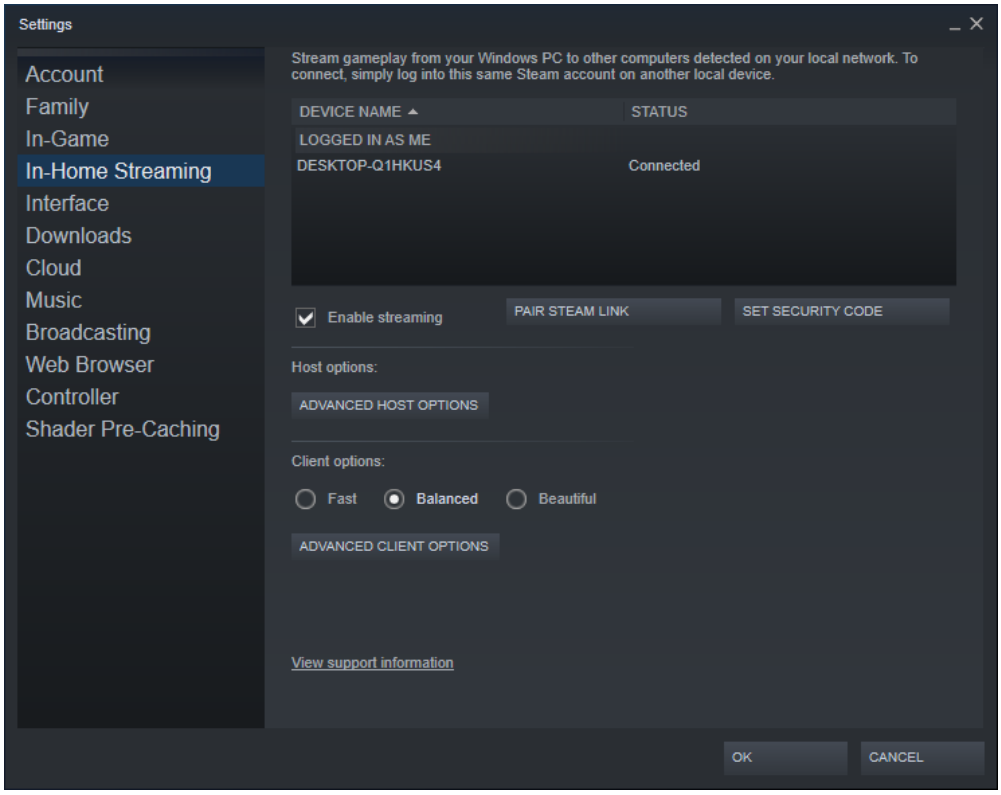

Figure 11. Steam In-Home Streaming settings

At this point the server is ready to stream the games. The games can be installed and started from the client machine. The testing of the service is done in the following section.

5.4 Testing

In order to start testing the service, both machines should be set up and configured. The client and the server need to be connected to the same network. In case of this project, they are connected to ZeroTier One virtual private network that was mentioned in the previous section. Steam has to be running on both devices and Steam In-Home Streaming should be enabled.

As the performance may vary depending on the type of game, for the testing purposes three games of different types are used. The most popular game types are First Person Shooter (FPS), Role Playing Game (RPG), and Real Time Strategy (RTS). The following games were chosen for the testing:

- **Borderlands** is an open world action role-playing FPS game developed by Gearbox Software in 2009. Borderlands has a cel-shaded cartoon-style graphics.
- **Divinity: Original Sin** is a fantasy RPG by Larian Studios with a tactical turn-based combat. Divinity has a party consisting of two main and two additional characters. The user is able to control all of the characters during the battle.
- **Northgard** is a real time strategy game developed by Shiro Games. It allows the user to operate a settlement of Vikings and lead them to the prosperity.

The testing started with Borderlands (Figure 12). Considering that it is an FPS game, its response delay should be as low as possible. High latency may have a negative result for the player and cause a frustration with the game process. Therefore, it is important to have the response delay less than 100 milliseconds. As it is difficult to measure the exact response time, only streaming latency is considered in this testing.

Figure 12. Borderlands gameplay test

After playing Borderlands, the following statistics was obtained:

- Streaming Latency: 18.98ms input, 38.68ms display
- Ping time: 39.87ms

- Estimated bandwidth: 2Mbps
- Packet loss: 0.00% (0.00% frame loss)

This is the approximate result and may vary depending on the internet connection and other factors. This result is acceptable for gaming, even though there were moments when the response delay was considerably high.

The next game tested was Divinity: Original Sin (Figure 13). Being an RPG, it doesn't need the response time to be as short as in case of FPS games. The maximum response delay allowed for RPGs is 500 milliseconds. That happens because in role playing games the playable character doesn't perform instant actions. Instead, the animation happens before the action is executed. This allows having longer delay without degrading user experience.

Figure 13. Divinity: Original Sin gameplay testing

The following statistics was obtained after playing Divinity: Original Sin:

- Streaming Latency: 22.71ms input, 41.67ms display
- Ping time: 38.22ms
- Estimated bandwidth: 19Mbps
- Packet loss: 0.00% (0.00% frame loss)

The latency is a little higher this time. The reason for this can be the better quality of graphics in the game and the higher screen resolution. Despite that, the

gameplay was satisfactory. The actions were smooth, and the game didn’t have any drastic interruptions.

The last test was done with the game Northgard (Figure 14). It is a strategy game and allows a response delay of 1,000 milliseconds. In case of RTS games, high latency is not always a problem. The player controls multiple game instances at the same time, which may leave the response delays unnoticed.

Figure 14. Northgard gameplay test

Statistics of the Northgard gameplay is the following:

- Streaming Latency: 22.99ms input, 46.14ms display
- Ping time: 43.67ms
- Estimated bandwidth: 17Mbps
- Packet loss: 0.00% (0.20% frame loss)

The latency is close to the one obtained from Divinity: Original Sin test. It is a good result for an RTS game. These two games have comparable graphics quality and the same screen definition, which explains similar results. There were moments when the latency increased, but it didn’t affect the gaming experience. Overall, the results can be considered as positive. The games were successfully running on the cloud server and could be streamed from it to the client PC.

There were minor problems with high latency and short interruption. Having faster internet connection can solve this issue. Additionally, having the cloud server location closer to the client machine will make the ping shorter and improve the response time. In this project the server is located in Frankfurt, Germany, while the client is in Mikkeli, Finland. Unfortunately, the closest server location that AWS provides, which is Stockholm, Sweden, doesn't provide necessary GPU instances. Therefore, these results were the best possible outcome that could be received in the current situation.

5.5 Instance optimization

Now that the service is working as expected, there is a need for optimizing the instance. Considering, that AWS EC2 instances are charging the user per hour of usage it is important to turn the instance off, when it is not used. In case of On-Demand Instances this is not a problem. On-Demand Instances are owned by the user and can be shut down or hibernated at any desired moment. In this project, however, the Spot Instance is used. In case of Spot Instances, the user is provided with spare computational power. In other words, this means that the user does not technically own the computational resources and is not able to shut down the instance manually. This causes a problem in paying for an idle server, as the instance is running without interruption. There is a way to solve this issue. The solution is to create an AMI (Amazon Machine Image) of the configured server.

As mentioned in the section 5.2, AWS provides its own AMIs that allow the user to start the work instantly without installing the operating system or the basic software. It is possible, however, to create a custom AMI that saves the state of the machine (Figure 15). Having the image of the instance allows canceling the Spot Request and to terminate the instance.

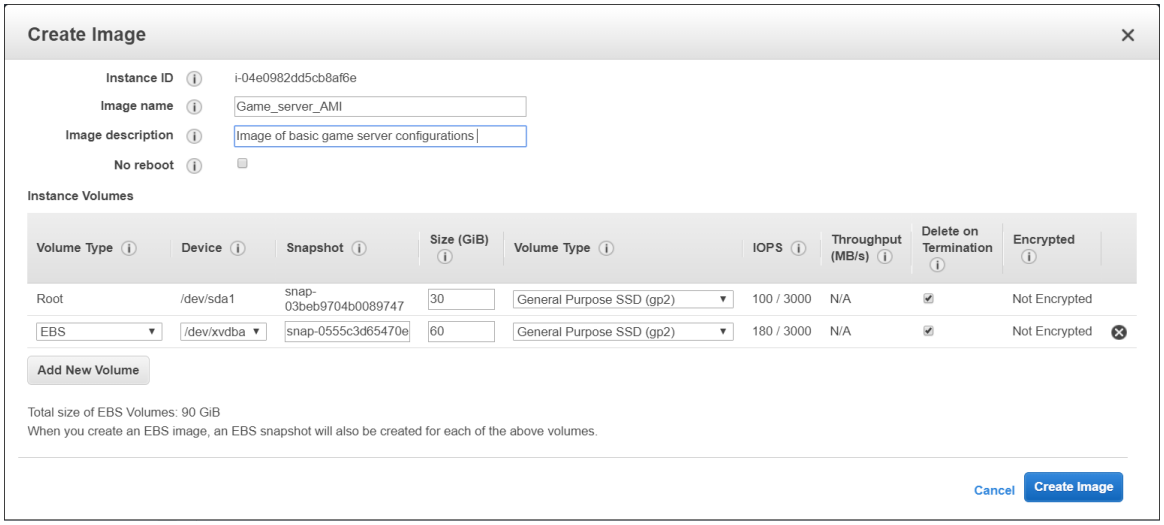

Figure 15. Creating instance AMI

Figure 15 shows the process of creating the image from the existing instance. The snapshots of the volumes are made to preserve the contents. After the AMI is created, the instance can be terminated without a worry of loosing any configurations. The next time the server needs to be started, a new Spot Request should be made. This time, instead of the AWS premade AMI, the custom AMI should be used.

Another thing that can be done to optimize the instance creation is creating a launch template. This way the instance can be created considerably faster than from scratch. AWS allows creating a launch template with all the necessary configurations. As Figure 16 shows, the user can specify the AMI for the instance, the instance type, the key pair name, the network type and the security group. All the configurations can be either specified or left empty. In the second case, they need to be defined later during the Spot Request creation.

Figure 16. Launch template settings

Now, Spot Instance can be terminated every time the server is not needed or used. It can be started again at any point by configuring new Spot Request from the launch template. The creation of Spot Request from a template takes less time and the instance doesn't have to be running uninterrupted anymore. This way a significant amount of money is saved.

6 CONCLUSIONS

The purpose of the study was to learn about cloud gaming and cloud computing, to implement a private cloud gaming service by using Amazon Web Services and to test the final result. Through the process of the study a broader understanding of cloud gaming technology was obtained. The study showed that the implementation of the service requires certain knowledge of cloud computing, networking, server administration and other skills. A study of cloud services was made. All the advantages and disadvantages were researched and considered in the process. During the implementation of the project the AWS instance was set up and configured successfully and the necessary software was installed on the server. In order to obtain the result of the project, gameplay tests were made.

The theoretical part helped to understand cloud gaming and cloud computing in more depth. It showed the importance of these technologies. In the process the benefits and challenges were explored, and the understanding of the architecture

was gained. The two biggest cloud service providers like Amazon Web Services and Azure were investigated.

The conclusions made during the implementation of the cloud gaming service were that cloud gaming is a growing technology and needs more time to develop. There are still some challenges currently. The Internet connection speed is very important for a positive gaming experience. Any interruptions in the connection can cause a user frustration. During the testing it became clear that minor interruptions are not an issue for some types of games. However, for some games, such as first player shooters, the fast response time is extremely important. Therefore, before the performance issue is solved, users will not prefer cloud gaming to the classical gaming approach.

REFERENCES

AWS. 2019. AWS Documentation. WWW document. Available at: https://docs.aws.amazon.com/index.html [Accessed 16 April 2019]

AWS. 2019. About AWS. WWW document. Available at: https://aws.amazon.com/about-aws/ [Accessed 15 April 2019]

Bozicevic, V. 2018. Cloud computing benefits: 7 key advantages for your business. WWW document. Available at: https://www.globaldots.com/cloud-computing-benefits/ [Accessed 21 March 2019]

Cai, W., Shea, R., Huang, Ch.-Y., Chen, K.-T., Liu, J., Leung V. C. M., Hsu, Ch.-H. 2016. A survey on cloud gaming: future of computer games. PDF document. Available at: https://www.researchgate.net/publication/306006176_A_Survey_on_Cloud_Gaming_Future_of_Computer_Games [Accessed 9 April 2019]

Cai, W., Shea, R., Huang, Ch.-Y., Chen, K.-T., Liu, J., Leung V. C. M., Hsu, Ch.-H. 2016. The future of cloud gaming. WWW document. Available at: https://www.researchgate.net/publication/298799626_The_Future_of_Cloud_Gaming [Accessed 27 March 2019]

D'Argenio, A. M. 2018. The past and future of cloud gaming: will it ever work. WWW document. Available at: https://www.gamecrate.com/past-and-future-cloud-gaming-will-it-ever-work/21044 [Accessed 21 March 2019]

Huang, Ch.-Y., Hsu, Ch.-H., Chang, Y.-Ch. & Chen, K.-T. 2013. GamingAnywhere: An open cloud gaming system. PDF document. Available at: http://dirl.iis.sinica.edu.tw/pub/huang13_gaming_anywhere.pdf [Accessed 27 March 2019]

Huth, A., Cebula, J. 2011. The basics of cloud computing. PDF document. Available at: https://www.us-cert.gov/sites/default/files/publications/CloudComputingHuthCebula.pdf [Accessed 10 March 2019]

Hwang, K., Fox, G. C., Dongarra, J. J. 2012. Distributed and cloud computing. Waltham: Elsevier, Inc.

Microsoft Azure. 2019. Overview. WWW document. Available at: https://azure.microsoft.com/en-us/overview/?v=48092-1914 [Accessed 17 April 2019]

Murugesan, S. & Bojanova, I. 2016. Encyclopedia on cloud computing. 1st edition. Chichester, West Sussex: John Wiley & Sons, Incorporated.

Perlman, S. 2010. OnLive: Coming to a screen near you. WWW document. Available at: https://web.archive.org/web/20100312043136/http://blog.onlive.com/2010/03/10/onlive-coming-to-a-screen-near-you/ [Accessed 27 March 2019]

Perry, D. 2011. Gaikai is live. WWW document. Available at: https://dperry.com/2011/02/06/gaikai_is_live/ [Accessed 27 March 2019]

Semsarzadeh, M., Hemmati, M., Javadtalab, A., Yassine, A. & Shirmohammadi, Sh. 2014. A Video Encoding Speed-up Architecture for Cloud Gaming. PDF document. Available at: https://www.researchgate.net/publication/268389507_A_Video_Encoding_Speed-up_Architecture_for_Cloud_Gaming [Accessed 9 April 2019]

Shea, R., Liu, J., Ngai, E. C.-H., Cui, Y. 2013. Cloud gaming: architecture and performance. PDF document. Available at: http://www.cs.sfu.ca/~jcliu/Papers/CloudGaming.pdf [Accessed 27 March 2019]

Stadia. 2019. WWW document. Available at: https://store.google.com/us/magazine/stadia?hl=en-US [Accessed 27 March 2019]

Techopedia. 2019. Pay as you go (PAYG). WWW document. Available at: https://www.techopedia.com/definition/26951/pay-as-you-go-payg [Accessed 15 April 2019]

Techopedia. 2019. Cloud gaming. WWW document. Available at: https://www.techopedia.com/definition/26527/cloud-gaming [Accessed 19 March 2019]

Torry Harris. 2019. Cloud computing overview. PDF document. Available at: https://www.thbs.com/downloads/Cloud-Computing-Overview.pdf [Accessed 20 March 2019]

ZeroTier. 2019. WWW document. Available at: https://www.zerotier.com/ [Accessed 19 April 2019]