Tampere University

Saana Saarteinen

CLOUD COST OPTIMIZATION & CAPACITY MANAGEMENT

Master of Science Thesis
Faculty of Engineering & Natural Sciences
Henri Pirkkalainen
Samuli Pekkola
May 2020

ABSTRACT

Saana Saarteinen: Cloud Cost Optimization & Capacity Management
Master of Science Thesis
Tampere University
Master's Degree Program in Information & Knowledge Management
May 2020

Public cloud services have recently gained immense popularity. Public clouds offer several service model options that support varying business needs. Current and future technological trends and many technical benefits make cloud environments a great solution for organizations. However, alongside the increasing use of cloud services, cost related issues have become evident. Organizations from varying industries are facing higher costs than expected. Failing to take cost optimization and capacity management into consideration has resulted in rising costs.

The objective of this thesis was to study how public cloud cost optimization and capacity management can form an effective business process that tackles the current cost related issues with cloud computing. This thesis was conducted as a qualitative case study for an international industrial company. The Process-Oriented Knowledge Management (PKM) framework was used to model the process and include pivotal cost optimization and capacity management activities.

The result of this thesis is a business process that takes cost optimization and capacity management activities into account for IaaS and PaaS service models. The business process was created from a cloud consumer point of view and includes the planning and run phases of an applications cloud journey. Pivotal activities, instruments, tools, knowledge, roles and responsibilities were identified and included within and along the process to ensure the ongoing development and accuracy of cost optimization and capacity management in organizations.

Keywords: Public Cloud, Cost Optimization, Capacity Management

TIIVISTELMÄ

Saana Saarteinen: Pilvipalveluiden kustannus optimointi ja kapasiteetin hallinta
Diplomityö
Tampereen yliopisto
Tietojohtamisen DI-tutkinto-ohjelma
Toukokuu 2020

Pilvipalvelut ovat nykyisin hyvin suosittuja. Ne tarjoavat useita palvelumalleja, jotka tukevat liiketoiminnan nykyisiä ja tulevia tarpeita. Pilviympäristö on hyvä ratkaisumalli monille organisaatioille, koska käyttämällä pilvipalveluja organisaatiot pystyvät hyödyntämään uusia teknologisia kehityssuuntia. Pilvipalvelujen käytön kasvaessa kustannukset ovat nousseet yhä tärkeämmäksi tekijäksi. Kustannustason nousu on tullut osittain yllätyksenä monilla toimialoilla. Pilvipalvelujen kustannusoptimoinnin ja kapasiteetin hallinnan puute on osa syytä kustannustason hallitsemattomaan nousuun.

Tämän työn tavoitteena on ollut tutkia, miten toimivalla liiketoiminnan prosessilla voidaan optimoida pilvipalvelujen kustannuksia ja kapasiteettia. Työ toteutettiin kvalitatiivisena tapaustutkimuksena kansainväliselle teollisuusyritykselle. Prosessin mallintamiseen käytettiin PKM -viitekehystä, jonka avulla koottiin prosessin keskeiset aktiviteetit kustannusoptimoinnin ja kapasiteetin hallinnan osa-alueilta.

Diplomityön lopputuloksena syntyi liiketoimintaprosessi, joka määrittää keskeiset aktiviteetit IaaS ja PaaS -palvelumallien kustannusten optimoinnille ja kapasiteetin hallinnalle. Prosessi on luotu pilvipalvelujen käyttäjälle. Se sisältää sovelluksen osalta pilvipalvelujen käytön suunnitteluvaiheen sekä pilvipalvelun elinkaaren aikaisen vaiheen. Osana prosessia määritettiin prosessin eri vaiheiden keskeiset aktiviteetit, välineet, työkalut, tieto, roolit ja vastuut. Näiden avulla varmistetaan kustannusten optimointi ja kapasiteetin hallinnan toimivuus ja sen jatkuva kehitys organisaatiossa.

Avainsanat: pilvipalvelut, kustannusten optimointi, kapasiteetin hallinta

PREFACE

This master's thesis project has been a very interesting and educational journey. I am truly grateful for the opportunity that was given to me by the case company. I would like to thank my thesis supervisor for the interesting and relevant topic, and for the great support and guidance during this project. I would also like to express my gratitude to all the individuals that took part in the interviews.

I would also like to thank my university examiner Henri Pirkkalainen for the valuable feedback and guidance during this thesis, and for always responding promptly with insightful advice.

Most importantly, I would like to thank my parents for always believing in me and supporting me throughout the course of my studies. In addition, I would like to thank my mother for all the comments and feedback towards the end of this thesis.

Tampere, 19.05.2020

Saana Saarteinen

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

API	Application Programming Interface
BYOL	Bring Your Own License
CAPEX	Capital Expenditure
CPU	Central Processing Unit
ERP	Enterprise Resource Planning
IaaS	Infrastructure as a Service
IT	Information Technology
KM	Knowledge Management
OPEX	Operating Expense
PaaS	Platform as a Service
PAYG	Pay as You Go
PKM	Process-Oriented Knowledge Management
RAM	Random Access Memory
ROI	Return on Investment
SaaS	Software as a Service
SW	Software
TCO	Total Cost of Ownership
VM	Virtual Machine

1. INTRODUCTION

Cloud computing has become well known across industries and continues to gain a wider customer base. Cloud services revenue has been predicted to reach $200 billion in 2020 and continues to displace traditional on-premises investment options. The cloud has become an increasingly valuable solution for organizations, as it serves as a gateway to future Information Technology (IT) trends. Digitization, Application Programming Interfaces (API), Artificial Intelligence (AI) and the Internet of Things (IoT) are some of many of the trends that push businesses towards cloud solutions. (Ward & Slattery 2018) Furthermore, elasticity, scalability, reduced investment and operating costs, as well as increased flexibility have been recognized as pivotal factors that lead to cloud adoption across industries (Maresova, Sobeslav & Krejcar 2017).

Although cloud computing enables consumers to select deployment and service models that support business needs (Kavis 2014) and has the ability to transform industries with current and future technological advancements (Ward & Slattery 2018), costs may quickly become an issue (Loten 2018). Disregarding the essential fact that cost optimization as well as capacity management are ongoing activities, will lead to a faulty cloud adoption process, resulting in costs that are higher than expected (Amazon 2018). Without sufficient cost optimization and capacity management in place, and falling into the trap of overestimating cloud capacity, costs regardless of the chosen deployment and service model will rise uncontrollably, resulting in IT budget losses that can accumulate to millions of dollars. (Loten 2018)

Organizations with applications in the cloud and currently shifting applications to a cloud environment are facing higher costs than expected (Loten 2018). Research has been conducted on cost optimization and capacity management in the cloud however, existing research mainly focuses on both areas as separate entities. Capacity management processes are available for on-premises and cloud solutions however, majority of the processes are depicted from a cloud providers point of view. On the other hand, cost optimization processes specific to the cloud are not as common and are often focused on one area of cost optimization rather than viewing the activity from a process perspective. Several practice-based models are available that are specific to cost optimization prac-

tices in cloud environments. Cost optimization and capacity management are intertwined, which is why it is important to understand how they affect and complement each other and form a process to avoid unnecessary costs.

The objective of this thesis is to study how public cloud cost optimization and capacity management can form an effective business process that tackles the current cost related issues with cloud computing.

RQ1: How can effective cloud cost optimization and capacity management support the optimization of cloud costs?

RQ2: How to design business processes to account for cost optimization and capacity management?

The research questions will be answered based on a literature review and an empirical study. Research question number two will be used to assist in the formulation of the business process, which will combine relevant information gathered from research question one. The goal is to create a process that focuses on essential cost optimization and capacity management areas during the preparation and planning phase prior to moving applications to a cloud environment, and for applications that are already deployed in a cloud environment. The Process-Oriented Knowledge Management (PKM) framework was chosen to design the cost optimization and capacity management process, as it specifically focuses on knowledge management and processes within organizations, while keeping business value in mind.

This thesis was conducted as a case study for an international industrial company. The literature review includes two different chapters, cloud cost optimization and cloud capacity management. The methodology chapter describes how the research was conducted. Chapter five presents the empirical results. The discussion chapter further combines key findings from the literature review and empirical study, as well as presents the business process. Furthermore, chapter seven concludes this thesis with detailing how the objectives of the thesis were met, contributions, limitations and suggestions for future research.

2. CLOUD COST OPTIMIZATION

Savings, ease of management and scalability are strongly associated with the term cloud computing and have been claimed to be the advantages over a traditional on-premise solution (Tak, Urgaonkar & Sivasubramaniam 2013). Chang, Walters & Wills (2013) similarly identify how cloud computing has enabled cost savings, agility and new business opportunities, as well as transformed the way organizations work. Furthermore, Lněnička (2013) states how the cloud increases scale of operations, while decreasing the cost of infrastructure. Therefore, the agile and dynamic cloud environment enables the rapid creation of services without any initial investments in hardware (Hähnle & Johnsen 2015). However, contrary to Tak et al. (2013) statement regarding the savings potential of cloud computing, Loten (2018) identifies the risks associated with rising costs.

Migrating applications to a cloud environment requires careful planning and taking various factors into consideration (De Capitani Di Vimercati, Foresti, Livraga, Piuri & Samarati, (2013). Preimesberger (2017) suggests establishing a business case before making the decision to move from an on-premise environment to the cloud. Mithani, Salsburg & Rao (2010) similarly state that prior to any business workload shifts from an on-premise environment to a cloud environment, the shifting of workloads must ensure and justify benefits to the business. The overall cost benefit is a pivotal factor when making a business case (Preimesberger 2017).

Evaluating and comparing potential cloud service plans and having the ability to match the appropriate plan with business needs has been identified as a challenging task. Therefore, understanding the feasible and possible options on the market is important prior to moving applications to the cloud, especially with the increasing demand for utilizing cloud services. (De Capitani et al. 2013) Furthermore, awareness on costs associated with migrating applications to the cloud play a vital role on the size of the cloud bill. In addition to cost awareness prior to the cloud, maintenance and support are ongoing activities, even when applications have already been migrated to the cloud. (Preimesberger 2017)

2.1 Cost Optimization

Cost optimization is strategic by nature and typically portrayed as programmatic. Cost optimization aims to create structured improvements, focusing on long-term achievements. (Ganly & Naegle 2019) Cristea (2017) introduces an economic model which can be used to identify the different stages of cost optimization.

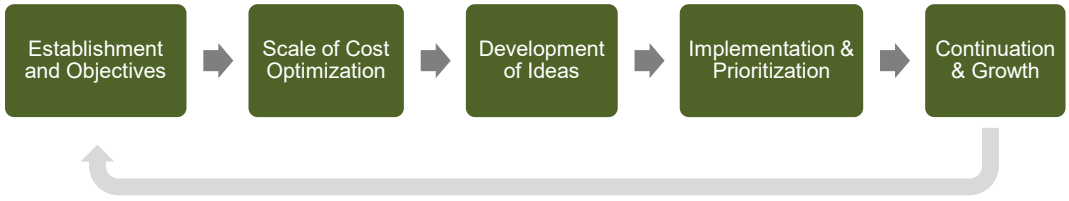


Figure 1. *Economic cost optimization model (adapted from Cristea 2017 & KPMG 2008)*

Establishment and objectives in figure 1 depict the beginning of the process. This requires a clear understanding of the objectives and ways to establish a cost optimization program. The second phase studies the scale of the cost optimization. The development of ideas phase in figure 1 analyzes the available opportunities for cost optimization and investigates which opportunities should be developed. Moreover, the implementation and prioritization phase select the appropriate initiatives to be implemented and establishes a prioritized implementation order. The continuation and growth phase in figure 1 analyze possible improvement areas as an ongoing activity, in order to maintain the benefits of cost optimization. (Cristea 2017)

In addition to the economic model, Cristea (2017) depicts a model which uses an innovative strategy in the formulation of cost optimization initiatives.

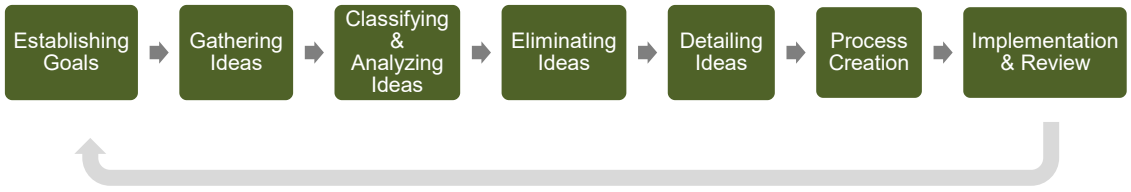


Figure 2. *Innovative strategies & cost optimization (adapted from Cristea 2017 & Khoury 2010)*

Cristea (2017) highlights how any strategy with innovative cost optimization components should focus on establishing goals. Effort must be placed in collecting innovative ideas from staff, with the use of appropriate mechanisms. The gathered ideas should then be filtered appropriately, eliminating ideas which are not feasible, and further detailing the ideas that are classified as feasible. The development of the implementation plan and methods should be formed into a process. Monitoring and review should take place as the final stage of the innovative strategy for optimizing costs. (Cristea 2017)

Cancila (2015) introduces a practice-based framework for public cloud cost management. The framework assists organizations with tracking, budgeting and optimizing cloud spend (Cancila 2015).

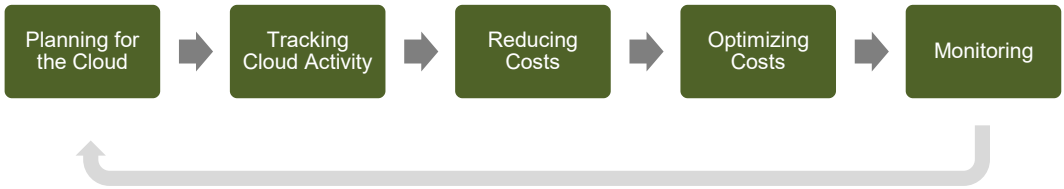


Figure 3. *Public cloud cost management framework (adapted from Cancila 2015)*

Cancila's (2015) practice-based framework studies cost optimization at a more detailed level specific to cloud environments, in comparison to Cristea's (2017) models. The planning for the cloud phase in Cancila's (2015) practice-based framework focuses on creating a forecast for the cloud spend. The tracking of cloud activity stage aims at attaining appropriate visibility to the cloud spend. The reduction of costs phase studies cost optimal deployment options within a public cloud environment. Moreover, optimization of costs focuses on using analytics to gain insight into the cloud environment. (Cancila 2015) The final stage of Cancila's (2015) practice-based framework taps into managing spend and processes and emphasizes the importance of forming a continuous process of the cost optimization initiatives.

An appropriate team for the execution of the cost optimization initiatives should be established (Cristea 2017). Ganly & Naegle (2019) identify how organizations often lack interest in optimizing costs during positive financial periods. Time as well as resources with the ability to perform cost optimization practices may also be limited. Both reasons can be categorized as risks which lead to disregarding cost optimization in organizations. However, when implemented and operated in a smooth manner at enterprise and functional levels, cost optimization can form innovative investments as a result of sustainably reinvested IT funds. (Ganly & Naegle 2019)

Cristea (2017), Cancila (2015) and Ganly and Naegle (2019) all depict a similar approach in the final stage of the cost optimization process. Cost optimization initiatives should be adapted as a continuous practice. The process should form a cycle that becomes a way of working within an organization. (Cristea 2017, Cancila 2015 & Ganly & Naegle 2019)

In addition, cost optimization requires the prioritization of various elements (Cristea 2017). According to Cristea (2017), these elements include:

- Potential monetary benefits
- Length of time it takes to implement the cost optimization initiatives

- Volume of resources that are necessary to implement the optimization decisions
- Risks accumulated alongside changes

Cost optimization initiatives should only be taken into consideration if most of the answers to the different prioritization elements listed on the left fall into the high priority column of figure 4 (Cristea 2017).

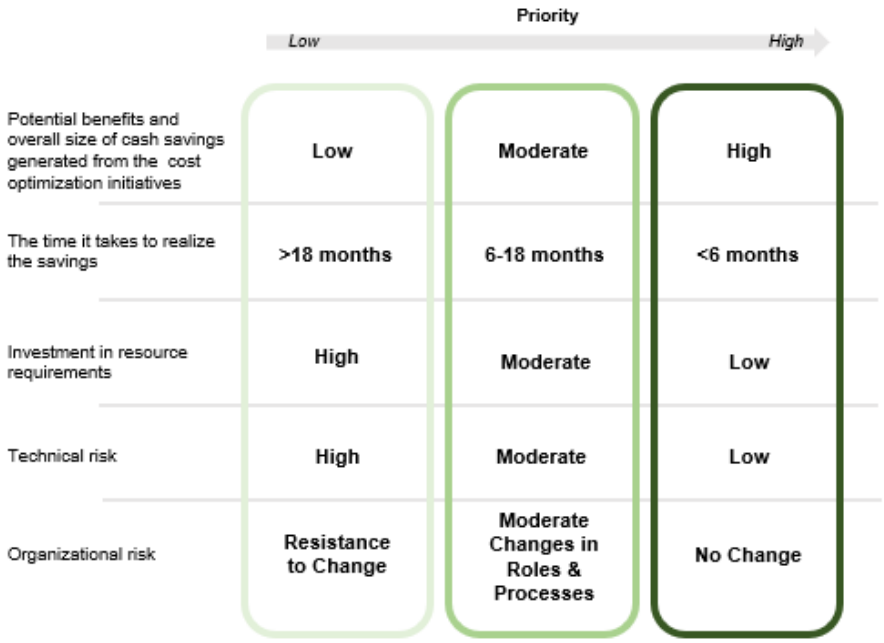


Figure 4. *Prioritizing cost optimization initiatives (adapted from Cristea 2017 & Gomolski & Kost 2009)*

2.2 Preparing for the Cloud

Organizations are eager to shift their business workloads from on-premise data centers to public clouds and gain the relevant cloud computing benefits. However, thorough analysis is necessary before moving workloads away from on-premise environments, as public clouds are extensively complex by nature. (Mithani et al. 2010) The migration of applications from an on-premise to a cloud environment can be considered a strategic organizational decision (Alkhalil, Sahandi, & John 2017). Application migration includes the shifting of an application from an on-premise to a cloud environment (Tran, Keung, Liu & Fekete 2011), to one of the cloud service models, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service (SaaS) (Jennings & Stadler 2015). The decision to migrate applications to the cloud has proven to be a rather difficult one, as a wide range of both technical and organizational aspects require in-depth evaluation (Alkhalil et al. 2017). Furthermore, finding an optimal deployment model that is suitable with the application requirements (Evangelinou, Ciavotta, Ardagna, Kopanel, Kousiouris

& Varvarigou 2018), as well as matching the application with the most cost-effective deployment model, cannot be categorized as a trivial task **(**Huang, Yi, Song, Yang, & Zhang 2014).

In order to reap the benefits of the cloud environment, applications must function properly in the cloud. This requires a clear understanding of the application at hand, and the cloud environment chosen for the deployment of the application. (Tran et al. 2011) The extent of software system complexity joined with a wide range of services and prices force consumers to evaluate a growing number of design alternatives (Koziolek, Koziolek & Reussner 2011) while keeping costs at a minimum. (Evangelinou et al. 2018) Furthermore, being ready for the cloud requires training. Consumers must have a clear understanding of the applications system environment, specifications and configurations. (Tran et al. 2011) As there are many cloud providers on the market it is important to examine the diversity of the cloud providers and the available technology stacks of cloud services (Evangelinou et al. 2018). Furthermore, licensing costs are incurred alongside scaling resources in the cloud (Suleiman, Sakr, Jeffery & Liu 2012). Having a good understanding of the cloud providers on the market and the offerings and technologies utilized assists in getting ready for a successful migration process (Tran et al. 2011).

A cost-benefit analysis should be included alongside the migration of an application to a cloud environment. This is an essential tool in assisting IT managers with identifying whether IT investment costs are outweighed by the benefits. (Tran et al. 2011) Furthermore, it is highly important to know how to manage dynamic computational resources of an application in a cloud environment, and especially focus on the trade-off between the amount of these computational resources and the costs (Andrikopoulos, Binz, Leymann & Strauch 2013). System designers must investigate a large array of alternatives and need to have the ability to evaluate costs, as the number of solutions is immense and application dynamics and performance tend to affect the costs (Evangelinou et al. 2018).

Many questions arise prior to migrating applications to a cloud environment. These include contemplation on what parts of the application to migrate, how to align and adapt the application to function in a cloud environment and if it would in fact be more beneficial cost wise to migrate the whole application. In order to tackle these dilemmas a clear understanding of the application and how it should be adapted to the cloud is necessary. (Andrikopoulos et al. 2013) All the essential pre-requisites must be thoroughly examined and documented by business and technology organizations prior to moving any business workloads into a cloud environment (Mithani et al. 2010). However, understanding the application behavior on cloud platforms prior to moving to a cloud environment is a key challenge for consumers. This is especially apparent when trying to determine the most

suitable environment to host application components from a cost point of view. (Evangelinou et al. 2018)

In addition, the impact of cloud adoption on the applications usual operations requires analysis (Andrikopoulos et al. 2013). It is crucial for businesses to understand the effect of different cloud environments on business processes (Lněnička 2013). The migration itself will be smooth, if the preparation for migration activities has been done accordingly (Tran et al. 2011). On the other hand, application owners often lack knowledge and awareness of how the migrated application components use cloud computing resources. In some cases, runtime behavior and usage of resources may be unknown or mistakenly altered for certain application components, as structural changes might occur during the software migration activity. (Evangelinou et al. 2018)

Cloud solutions are scalable from small offices to large enterprises. In addition, good cloud solutions enable simple use and adaptation of cloud services. (Case Company 2019b) The optimal solution for migrating applications may depend on many factors such as application characteristics, workload and the required Quality of Service (QoS) (Evangelinou et al. 2018). However, a crucial factor that must be considered prior to moving workloads to a cloud environment is whether the existing workload can and should in fact be deployed in a cloud environment (Mithani et al. 2010). Mithani et al. (2010), identify the types of workloads which are typically moved to public cloud environments. These include highly elastic workloads, test and pre-production systems, contextual applications including email, software development environments, batch processing jobs with limited security requirements, isolated workloads without latency requirements, storage solutions, backup solutions and data intensive workloads (Mithani et al. 2010). Lněnička (2013) identifies applications that have little interaction with back-end systems, applications with exponential demand increases, business intelligence and data mining applications, as well as test and development applications as best fits for a cloud computing environment.

On the other hand, not every application is fit for a cloud environment (Andrikopoulos et al. 2013). There are certain workloads that are not equipped to be hosted on virtual servers. Examples of workloads that are not a good fit for cloud environments include legacy workloads and workloads that need to meet precise service level objectives. Furthermore, applications that require physical servers for hosting are limited to a few choices public cloud wise. It is important to keep in mind that maximum mobility in public cloud environments is a possibility for business workloads that suit virtual image formats. (Mithani et al. 2010)

Furthermore, the case company identifies certain scenarios where public cloud solutions are not appropriate. These include solutions that are highly sensitive to network latency, such as real time applications which require close data integrations to function properly. Moreover, performance may become an issue for old software, as modifications are required before being suitable for a public cloud environment. In addition, data security might become a problem, as certain regulations mandate data to be audited on-premise. Certain software terms and conditions may also restrict the deployment of applications in a public cloud environment. There may also be situations where using public cloud services could result in a risk of vendor lock-in, with expensive exit plans. (Case Company 2019b)

2.3 Cloud Service Models and Optimization

There are three main cloud service models, IaaS, PaaS and SaaS (Han 2011). Jennings & Stadler (2015) similarly identify that a public cloud environment typically comprises of the IaaS, PaaS and SaaS service models. Determining the difference between the models depends on the level of abstraction of the offered service (Jennings & Stadler 2015).

In the IaaS service model, the cloud provider manages the underlying physical cloud infrastructure, providing services through virtualization (Han 2011). IaaS provides software developers access to bare infrastructure for computing, storage and networking (Louridas 2010). Amazon Elastic Compute Cloud (EC2) is an example of an IaaS service (Muhic & Bengtsson 2019).

In the PaaS service model, the cloud provider manages every layer in the service model stack, except the application layer (Han 2011). Software developers are given access to a development platform for designing, building, testing and deploying their own custom applications (Louridas 2010 & Muhic & Bengtsson 2019). Microsoft Azure's integrated environments (Muhic & Bengtsson 2019) and Microsoft SQL databases as a service are examples of PaaS services (Case Company 2019b).

In the SaaS model, cloud providers manage all the cloud infrastructure including the applications and application logic. This model enables end users to access applications through thin client interfaces i.e. web browsers. (Han 2011, Mell & Grance 2010 & Case Company 2019b) Examples range from standard email and office applications to more complex Enterprise Resource Planning (ERP) systems (Muhic & Bengtsson 2019).

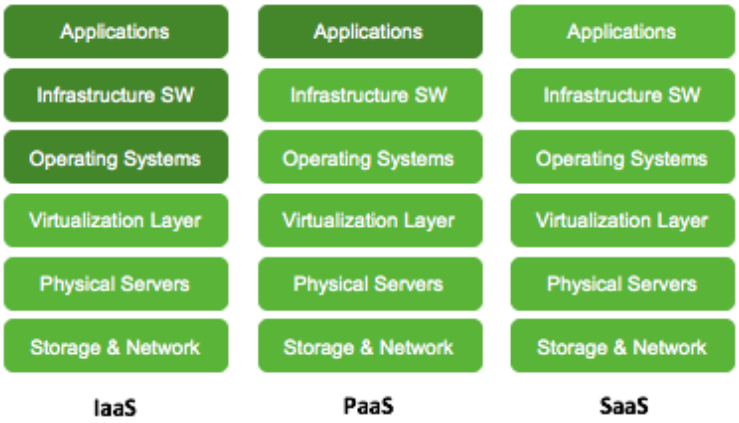


Figure 5. *Cloud service models (adapted from Rountree & Castrillo 2014)*

Consumers must evaluate and understand the complexities of the various service models (Sabharwal & Wali 2013). When planning and designing for migration, Microsoft highlights the importance of focusing on costs to ensure long-term success (Microsoft Azure 2018). Gartner splits the service models into five separate scenarios (Clayton 2018):

1. Rehost (“lift and shift”)
2. Revise
3. Rearchitect
4. Rebuild
5. Replace

The first and second scenarios, rehost and revise, are typically covered by the IaaS cloud service model. The rehosting (“lift and shift”) scenario entails the migration of virtual machines and data to the cloud IaaS. (Anderson 2018) This scenario avoids alterations to the systems. However, certain modifications are required to adapt to the new hosting environment. This scenario does not support cloud-native features. The revise scenario on the other hand, enables consumers to modify applications so that they can begin to utilize the advantages of cloud capabilities. These include elasticity, minimized resource usage and minimized operational overhead, by capitalizing on managed cloud services, such as database PaaS. In other words, consumers are given the option of optimizing the infrastructure and backing services of the application. This entails making minor changes to the code or leaving the code untouched, while reconfiguring the application, system and application dependencies. (Clayton 2018) Overall, this migration scenario does not yield major cost savings but is a fairly simple form of migration. Optimization is possible and goes hand in hand with resource usage and elasticity. (Anderson 2018)

The case company also suggests optimization activities specific to the IaaS service model (Case Company 2019b):

- Development phase optimization: Rightsizing capacity, autoscaling as a design and automating on-off capabilities for applications that do not require 24/7 uptime.
- Run/ production phase optimization: Monitoring capacity, reacting to and planning possible changes in capacity usage and opting for reserved instances when feasible.

The third and fourth scenarios, rearchitect and rebuild, belong under the PaaS cloud service model. The PaaS scenario entails migrating the application to the cloud middleware. (Anderson 2018) If artifacts of the application can be reused, the application is under constant rapid change, the application is either flexible or inflexible portability wise between cloud providers and there is time and an abundance of resources to rearchitect the application, then rearchitecting should be considered. However, if application portability to a cloud platform is considered difficult, existing artifacts cannot be reused, the application cannot be virtualized, there is no pressure time wise to get the application to the market, and resources and time are available to rebuild the application, then rebuilding the application may be the best option. (Clayton 2018) The potential cost savings of the PaaS migration scenario are high however, having the ability to implement cloud as a native application capability and leveraging the PaaS components is categorized as a difficult task. Optimization is possible by exploiting the elasticity features of PaaS deployments in cloud. (Anderson 2018) The case company identifies optimization possibilities for PaaS applications (Case Company 2019b):

- Development phase optimization: Designing the solution to scale, eliminating any extra and unnecessary capacity.
- Run/ production phase optimization: Data lifecycle management, identifying and removing orphaned resources and considering commitment possibilities.

The final scenario, replace, covers SaaS cloud service models. This scenario entails replacing a traditional application with a SaaS application. Replacing includes migrating all the users and data to the cloud and shutting down the application from an on-premise environment. (Anderson 2018) If a SaaS offering is available, and there is a possibility in investing in the SaaS option, then replacing should be considered (Clayton 2018). The cost savings potential of SaaS models falls somewhere in between the potential savings of IaaS and PaaS service models (Anderson 2018). The difficulty of a SaaS deployment is low. Optimization possibilities include users, entitlements (Anderson 2018) and data

(Case Company 2019b). The case company also suggest possible optimization activities for SaaS service models (Case Company 2019b):

- Development phase optimization: Sourcing and contracts with optimization requirements.
- Run/ production phase optimization: Optimizing users and usage.

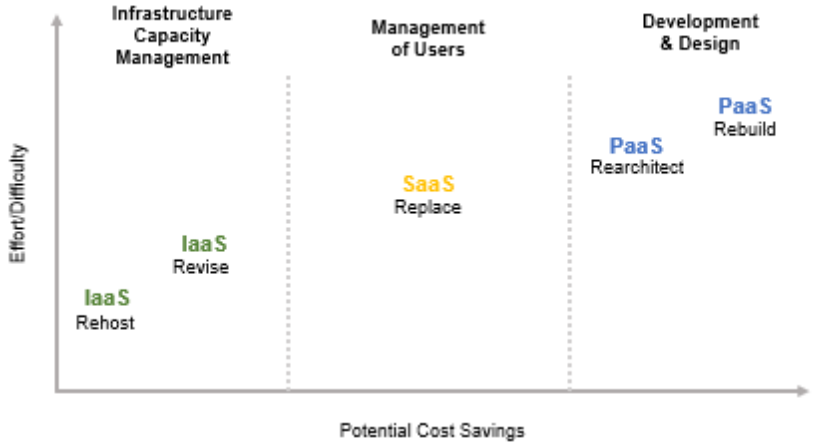


Figure 6. *Cost savings potential & difficulty of cloud service models (adapted from Case Company 2019b & Clayton 2018)*

In the case company, new IT solutions must primarily be considered as cloud-based solutions. Reasons for this include the fact that cloud solutions embody characteristics including fast deployment, evergreen models and scalable capacity and pricing. (Case Company 2019b) The case company (2019b) has a clear prioritization scheme regarding the different cloud service models:

1. The SaaS model must be considered first, as it yields best practice business processes outside of the core service. This service model may be used i.e. to fulfill a business process within an organization.
2. The PaaS model is suggested as a second choice. This service model enables rapid deployments with the possibility of digital differentiation.
3. Ultimately the IaaS service model should be considered. The IaaS service model enables users to gain elastic computing capacity.

Furthermore, Microsoft highlights that over time a migrated resource may shift to another type of workload. Reasons for this shift include changing business requirements, costs and usage. (Microsoft Azure 2018)

2.4 Cloud Cost Models and Optimization

Another area of the dynamic cloud environment that requires preparation is knowing which pricing models to consider and ultimately choose for deployment. Public cloud offering complexities make it difficult to understand the best strategy for movement not only in terms of technologies, but also in terms of complicated terminologies. (Mithani et al. 2010) For any type of migration, one area which affects the costs of migrating an application, particularly parts of an application to a cloud provider, are the pricing models offered (Andrikopoulos et al. 2013). Occasionally, the complexity of the available IaaS pricing models can make it harder to assess the actual monetary benefit of migrating applications to a public cloud environment (Jennings & Stadler 2015). Experts must have the ability to analyze and understand the pricing models and cloud offerings (Mithani et al. 2010).

The cloud continues to gain popularity as it has presented a clear case for reducing capital expenditure and turning it into operational costs (Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Patterson, Rabkin, Stoica & Zaharia 2009). Willcocks, Venters & Whitley (2013) identify how the pay as you go subscription-based model has enabled a shift in IT expenditure from capital expenditure to operational expenditure budgets. Jennings & Stadler (2015) similarly state how hosting applications in a cloud environment, such as the IaaS service model, lowers capital and operational expenses. The pay-per-use model has enabled the saving of fixed costs by allowing consumers to lease resources, instead of buying resources (Andrikopoulos et al. 2013).

Many argue that cloud computing can be considered cheaper in terms of Total Cost of Ownership (TCO) (Wu, Buyya & Ramamohanarao 2019). This however, is not a mutual opinion, as some believe cloud computing to not be cheap (Weinman 2012), as there is ambiguity behind the pricing models and the estimated build-up of real costs (Martens, Walterbusch & Teuteberg 2012). The varying pricing models are known to be overwhelming, especially as there are multiple cloud service providers on the market (Wu et al. 2019).

The reservation and on-demand plans are available for the disposal of cloud consumers (Chaisiri, Lee & Niyato 2009). The reserved pricing model ensures cloud resource certainty (Wu et al. 2019). Resource provisioning is generally cheaper when acquiring the reservation plan. However, contrary to the on-demand plan, the reservation plan must be obtained in advance. With the reservation model future demands may not be fully met, whereas the on-demand pricing model guarantees availability. (Chaisiri et al. 2009)

The on-demand model is a good fit for workloads with inconsistent consumption, as resources can be provisioned as needed and on an urgent basis (Singh & Chana 2015). On the other hand, the subscription model requires adequate knowledge on capacity management to ensure resources are aligned with the application needs, as this model provides workloads long-term reservations (Singh & Chana 2015). Workloads may also utilize a mix of these cost models (Wu et al. 2019).

Suleiman et al. (2012) identify four different pricing models including the subscription, per-use, prepaid per-use and the subscription and per-use model. Furthermore, Suleiman et al (2012) analyze workload patterns, economics of pricing models and elasticity of offerings to appropriately match workloads with the adequate pricing models.

The subscription model entails dedicated servers or reserved instances, which require a commitment. The commitment can be short-term or long-term and is often offered at discounted monthly/yearly rates. (Suleiman et al. 2012) Suleiman et al (2012) highlight how the subscription model is typically cheaper than the per-use model however, the application workload needs to be fixed and constant.

The per-use model, also known as the pay-as-you-go pricing model is used for on-demand servers. No commitment is required, and resources can be requested according to needs. On the other hand, the prepaid per-use model entails on-demand servers which are billed hourly from a prepaid credit without commitment requirements. Consumers must ensure that credit does not go below a certain limit as some providers may charge exceeding the limit on a per-use basis. However, the consumer must also pay attention to the unused credits, as refunds might not be possible with certain providers. Variable workloads with variable volumes go hand in hand with the per-use/ prepaid per-use model. These pricing models provide computing resources according to needs and prevent over or under provisioning scenarios. Therefore, workloads that are highly elastic and require resources on-demand to scale up and down should opt for the per-use/prepaid per-use pricing model. Furthermore, unpredictable workloads should capitalize on the prepaid per-use and per-use model combination as this enables very high elasticity for daily and hourly on-demand servers. (Suleiman et al. 2012)

The subscription and per-use model enable the renting of dedicated servers in advance, and the requesting of additional cloud servers on-demand that are billed according to the per-use cost model. This model combines the advantages of discounted dedicated servers for stable workloads, and the availability of on-demand instances for application workloads that fluctuate. In other words, fixed workloads with predictable spikes should combine the subscription and per-use pricing models. This assists in avoiding over or

under provisioning of the predictable spikes. By using this pricing model combination, high elasticity is available for the predictable spikes using hourly on-demand servers. (Suleiman et al. 2012)

Wu et al. (2019) identify seven different mainstream pricing models on the cloud market. These include the discount, reserve, on-demand, subscription, code on demand, bare metal and dedicated host pricing models (Wu et al. 2019). Figure 7 depicts five of these pricing models. The pricing model costs increase along the arrow in figure 7, with spot instances being the cheapest and code on demand, as well as on-demand pricing accumulating the highest costs. (Wu et al. 2019)

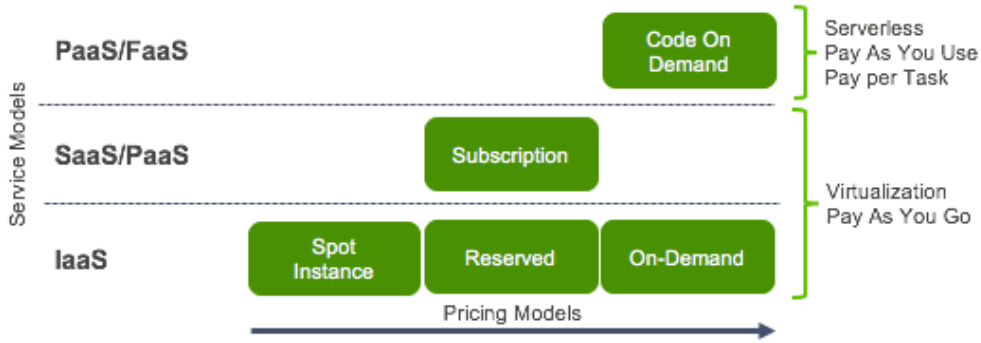


Figure 7. *Cloud service & pricing models (adapted from Wu et al. 2019)*

Sumalatha & Anbarasi (2019) on the other hand identify reserved instances as the cheapest pricing model, as demonstrated in figure 8, where the price is based on a static period of subscription. These resources are to be reserved in advance by consumers. It is important for consumers to understand the usage level of their resources in order to avoid overpaying for unused resources. (Sumalatha & Anbarasi 2019) Sumalatha & Anbarasi (2019) identify on-demand instances as the highest priced resources. The price remains constant and consumers pay according to usage. The price of the on-demand model does not fluctuate according to market demands. Spot instances on the other hand, allow consumers to specify the maximum amount they are willing to pay to run a particular instance type. This rate is usually lower than the on-demand rate. This pricing model goes hand in hand with the supply and demand of instances. In other words, spot instances are unused on-demand instances. The spot price will never exceed the maximum price specified by the consumer, and once price levels surpass the limit, the instance is automatically shut down by the cloud provider. (Sumalatha & Anbarasi 2019)

Figure 8. *IaaS pricing models (adapted from Sumalatha & Anbarasi 2019)*

Although the usage-based static pricing model remains the predominant business model for IaaS and PaaS providers, shifts towards dynamic pricing models have become apparent. Dynamic pricing entails lowering costs when the usage of cloud service providers' resources is low. By offering a dynamic pricing model, cloud service providers hope to attract greater levels of usage, which in turn increases resource usage and maximizes profits. Cloud consumers should analyze the possibility of utilizing these low-cost options for the resources they lease, in order to maximize profits. Cloud consumers may lack the ability to capitalize on the appropriate cost model and therefore require the assistance of a cloud broker. (Jennings & Stadler 2015)

Furthermore, a key challenge for consumers is how to select the most economical and elastic offering. The applications workload patterns and characteristics, as well as certain other factors influence the appropriate choice. It is important to keep in mind that there is no one-size-fits-all pricing model or offering type that would suit various application workload patterns. Achieving the most economical and elastic solution is a challenging task. (Suleiman et al. 2012)

Moreover, with various options for purchasing capacity, in order to optimize costs consumers must contemplate the following scenarios (Sabharwal & Wali 2013):

- The amount of capacity to be purchased upfront for a longer period to enable discounted pricing
- The amount of capacity that is needed on-demand
- The tactic with spot instances, to further enhance the use of lower cost pricing models
- The available SaaS options and comparing the cost of SaaS options to IaaS deployments

2.5 Cloud Governance

Providers and consumers have been the main stakeholders for on-premise solutions. The roles of the provider in an on-premise model include sales, installation, licensing, consulting and maintenance of the technology. The roles of the consumer include the use, owning, maintaining and upgrading of the on-premise systems. There is a clear shift in the roles of the relevant stakeholders in a cloud environment. Furthermore, new additional stakeholders become relevant alongside cloud adoption. (Marston, Li, Bandyopadhyay, Zhang & Ghalsasi 2011) It is crucial to include all the relevant stakeholders within an organization, when planning for a cost aware cloud adoption (Amazon 2018), as the cloud environment is very different from a traditional on-premise set up (Marston et al. 2011). Prasad, Green & Heales (2014) agree that including the relevant stakeholders is crucial for a successful cloud journey.

An organizations governance model must consider all the relevant stakeholders including external ones, such as the cloud service provider (Prasad et al. 2014). Prasad and Green (2015) suggest an end to end view on business and IT functional areas when utilizing the cloud, as interaction is needed between internal and external stakeholders (Prasad et al. 2014). Organizations, providers and providers partners will need to be more collaborative than before (Willcocks et al. 2013). Marston et al. (2011) further identify how Chief Information Officers (CIO) and Chief Technology Officers (CTO) need to work hand in hand to develop an appropriate cloud strategy for an organization. In addition, a smaller group of individuals should continually evaluate developments in the cloud from a cost perspective (Marston et al. 2011). It is also important to note that external stakeholders, such as public cloud providers business partners are well equipped to assist organizations in finding the best public cloud deployment options. However, for a public cloud providers business partner to ensure the smooth implementation and deployment of organizations business workloads to a cloud environment, the business partner needs to be aware of the organizations business processes. (Mithani et al. 2010)

Effective governance of the cloud services will result in many benefits including efficiency gains. The gained benefits will improve business processes. This in turn will enable reaching financial objectives and Return on Investment (ROI). (Peiris, Balachandran & Sharma 2010) Furthermore, an appropriate governance model will result in spending IT related money in a careful and well thought out manner. Proper management and governance of the cloud services in relation to an organizations business processes will assist in managing IT expenditure constrains. In other words, this will ascertain returns from IT investments within a reasonable time period. (Prasad et al. 2014)

Adopting cloud services requires constant alignment between service providers, service intermediaries and other relevant stakeholders. This continuous activity will ensure the use of cloud services in an efficient and justifiable manner. (Marston et al. 2011) Engaging the appropriate stakeholders positively effects business process performance, which in turn will lower the cost of operations (Prasad & Green 2015). To realize the benefits of the cloud, organizations need to develop appropriate competencies (Prasad & Green 2015). Instead of establishing completely new IT governance structures just for the cloud, organizations will most likely include the relevant qualities in their current IT governance structures to avoid unnecessary costs (Debreceny 2013).

Cloud governance should be split into three different levels, business, service and technical governance. Business related governance deals with cloud consumption and management. Service governance is related to the provider and includes, tracking, measuring, monitoring and enforcement of the cloud services. Technical governance relates to the more technical understanding of cloud services. (Prasad et al. 2014) Specific qualities need to be present in governance structures for appropriate management of cloud services, as competence of the cloud will lead to better use of the cloud, resulting in improved business IT-alignment and value (Prasad & Green 2015). Willcocks et al. (2013), similarly state how it is important that organizations pay attention to the skill sets and knowledge of their employees, as this will impact the adaptation of the cloud services.

Prasad et al. (2014) suggest a Chief Cloud Officer (CCO), a Cloud Management Committee (CMC), a cloud service facilitation center and a Cloud Relationship Center (CRC), as possible governance structures for cloud computing services, to ensure that cloud services match the organizations business processes and financial objectives. A CCO, either an individual or team would be experts in cloud services, covering some of the technical governance. Having in-house talent regarding cloud services is crucial. The alignment of the cloud and business processes within an organization will guarantee a more beneficial cloud journey. The CMC would combine different level stakeholders to oversee the adoption of cloud services. Stakeholders include members within the organization, cloud service providers and cloud service intermediaries. The cloud service facilitation center would overlook the operational management of the cloud services in organizations. (Prasad et al. 2014) This includes issue resolution, performance monitoring, and tactical decisions (Block 2012). The CRC would sit between the cloud service provider and the service users. The CRC would ensure policies are followed and that the objectives of the service are in line with the use of the service. (Prasad et al. 2014) As

there are multiple systems and applications in an IT environment which are run by different teams within an organization (Amazon 2018), cloud service policies play an immense role in the cloud (Prasad et al. 2014).

Amazon lists four relevant stakeholders. These include Chief Financial Officers (CFO), business unit owners, tech leads and third parties. The CFO and the organizations financial controllers are required to have a thorough understanding of the models of consumption, purchasing options as well as the monthly billing process and data that comes with the billing. CFOs and financial controllers must understand how the procurement processes, incentive tracking and financial statements may be affected. Business unit owners need proper understanding of the cloud business model. This is an essential role when forecasting growth and system usage is required. In addition, the business unit owners need to have a firm grip on the different purchasing options. Tech leads must have the ability to implement systems that achieve goals of the business. As an example, this includes translating cost factors into system attributes or adjustments. Furthermore, third parties must be aligned with the financial goals of the organization. Third parties tend to contribute towards reporting and analysis of systems that they manage. (Amazon 2018)

Microsoft emphasizes the importance of a cost-conscious organization. There are three activities which should be continuously performed by different parties within an organization. These activities include visibility, accountability and optimization. Visibility should enable cost consciousness. Consistent reporting should be available for teams that are utilizing cloud services, finance teams involved with budgeting, and management teams that take ownership of the costs. This requires the right type of reporting, good resource organization, an appropriate tagging strategy and proper access controls. Accountability includes the ability to have clear budgets for the cloud adoption efforts. Budgets need to be well established and communicated, as well as created based on realistic expectations. Optimization creates the cost reductions. Resource allocations are tweaked to reduce the cost of workloads in the cloud environment. Balance between cost reductions and performance requires the input of multiple parties. The optimization process is repetitive by nature and may require experimentation. A cloud strategy team, cloud adoption team, cloud governance team and cloud center of excellence should conduct the visibility, accountability and optimization activities. (Microsoft Azure 2019)

Microsoft highlights the importance of tagging and recommends it as an initial step towards proper governance of any environment (Microsoft Azure 2019b). Tags are used throughout industries as a useful way to organize resources (Malik, Chard & Foster

2014). Tags are also used as knowledge retrieval and information discovery tools (Matthews, Jones, Puzo, Moon, Tudhope, Golub & Lykke Nielsen 2010). In a cloud environment, tags can assist in organizing resources in a systematic manner that assists with tracking and raising awareness on resource consumption costs within an organization. Tracking consumption and costs should include the ability to match usage behavior with the correct user, system or defined entity. (Amazon 2018) Often used tags within organizations include business unit, department, billing code, geography, environment, project and workload (Microsoft Azure 2019b).

Sultan & van de Bunt-Kokhuis (2012) mention how future technological innovations could potentially have a profound effect on the way organizations conduct business. As a result, cultural issues are inevitable for organizations that use cloud computing services. Consumers must be prepared and willing to implement cultural changes, especially in the way they view their IT resources and infrastructure. (Sultan & van de Bunt-Kokhuis 2012) Organizations are known to develop their own unique cultural identity. The speed of cloud implementation will partially be determined by an organizations culture. (Willcocks et al. 2013)

2.6 Cloud Sourcing

Schneider & Sunyaev (2016) define cloud sourcing as an organization's decision to integrate cloud services from cloud providers into their own IT landscape. This entails an assessment of the potential cloud providers and their offerings, such as the different service models (IaaS, PaaS, SaaS) (Muhic & Johansson 2014). Cloud sourcing and cloud computing introduce a new form of organizational flexibility (Teece 2018). However, the shift from traditional IT-sourcing to cloud sourcing has proven to be a challenging proposition for larger firms (Willcocks et al. 2013), as cloud computing affects the sourcing processes of organizations (Muhic & Johansson 2014). Muhic & Johansson (2014) study the potential of cloud sourcing becoming the next generation of outsourcing. Traditional IT outsourcing and cloud computing have several similarities however, task responsibilities, advanced governance approaches, short term contracts based on usage, standardized services and the luxury of self-service procurement force organizations to rethink their sourcing processes (Schneider & Sunyaev 2016).

Lower costs, facilitated expansion, standardization of processes and more frequent maintenance of programs and systems are identified as the usual arguments for cloud sourcing in organizations (Muhic & Bengtsson 2019). Schneider & Sunyaev (2016) identify technological aspects and cost savings as the most determinant factors of cloud

sourcing decisions. Similarly, Muhic & Johansson (2014) find cost benefits a major motivator for sourcing cloud services and highlight how the flexible and elastic nature of cloud resources are the main advantages of cloud sourcing. Hayes (2010) also identifies how cloud sourcing has the potential to bring operational and cost related benefits. Therefore, greater flexibility and cost related benefits are pivotal in motivating the shift of applications to a cloud environment (Muhic & Bengtsson 2019). Additionally, Muhic & Johansson (2014) list access to talent as another factor that affects the motivation to source cloud services. However, the advantages of cloud sourcing, such as the on-demand and pay per use cost model, as well as the relief of managing IT-resources are not as easy to reap as it may seem (Willcocks et al. 2013).

Traditional IT-sourcing entails a one-to-one relationship between clients and vendors (Vithayathil 2018). Cloud sourcing on the other hand requires the ability to interact and manage an eco-system of cloud provider firms. Cloud provider firms include i.e. cloud brokers, cloud providers, cloud sub providers and IT-consultant firms. (Willcocks et al. 2013) The self-service nature of the cloud however, puts organizations in the role of the consumer, producer or co-producer of cloud services (Willcocks et al. 2013). Willcocks et al. (2013) identify how similarly to IT-outsourcing, distinctive in-house skills are required to ensure that cloud computing is used in an effective manner. Defining the computing requirements will need to be done specifically with an understanding of the cloud computing offerings (Willcocks et al. 2013).

Cloud sourcing often starts with technology-triggered processes. This entails attempting to make the cloud sourcing solution work as intended. However, achieving stability from a technical and operational standpoint does not eliminate business-oriented issues. For this reason, business opportunities must be a part of cloud sourcing. In addition, implementing development work and re-organization activities related to cloud sourcing have proven to be important. (Muhic & Bengtsson 2019) Strategic and business model changes should be included in the motives of cloud sourcing (Muhic & Bengtsson 2019), as cloud sourcing is closely related to an organization’s IT strategy (Muhic & Johansson 2014). Moreover, organizations that tap into the innovation possibilities of the cloud are bound to benefit from cloud computing at an even larger scale than solely focusing on the financial benefits. For this reason, limitations on reaching financial and innovative goals need to be identified and understood in order to reap the long-term benefits of the cloud. (Willcocks et al. 2013) Furthermore, it is important for organizations to identify when the cloud provider is not delivering services according to the agreement. In these cases, terminating the contract and finding a new cloud provider is essential. (Muhic & Bengtsson 2019)

2.7 Licensing in the Cloud

Cloud computing is altering the manner in which software is used, delivered and sold (Ojala 2013). Software licensing costs are apparent when migrating an application to the cloud. Whether performing a partial migration of some of the applications functions, migrating the entire software stack of the application, replacing components with cloud offerings or cloudifying the application, software licensing costs are incurred. (Andrikopoulos et al. 2013) Software licensing is considered a major obstacle when migrating applications to the cloud (Armbrust et al. 2009), and can be identified as a non-technical issue, that must not be overlooked (Reese 2009). Suleiman et al. (2012) state that the adding and removing of instances has been simplified for end users, leaving the consumer vulnerable to launch software applications on instances without having proper licensing in place, or reaching license thresholds such as maximum number of concurrent users/ Central Processing Units (CPU). Therefore, scaling systems in the cloud may lead to unintended license agreement violations (Andikopoulos 2013 & Reese 2009).

Vendors can sell software using combinations of different models ranging from server-based licensing to software renting (Ojala 2013). Traditional on-premise software licensing is typically based on the number of CPUs (Reese 2009 & Ojala 2013) or a consumer buying a single license for a single user or computer (Ojala 2013). This however, does not correlate with the dynamic nature of the cloud, in terms of number of instances and CPUs offered (Andrikopoulos et al. 2013). Application software needs to have the ability to scale up as well as down in a rapid manner. This type of software requires a pay-for-use licensing model, in order to align with the benefits of cloud computing. (Armbrust et al. 2009) In other words, the workload pattern of an application is a factor that effects license management (Suleiman et al. 2012).

The economic value of an application could be directly linked to software licensing issues. These include penalties, additional license fees, elasticity and restricted launching of servers. (Suleiman et al. 2012) Therefore, Suleiman et al. (2012) urge consumers to consider the following issues regarding software and system licensing in the cloud:

- The best application to workload fit license wise, out of all the cloud service provider license type offerings
- How licensing models on cloud server instances impact the economics and scalability of the application
- The ability to monitor and control various software and system licenses on all running server instances

Mohan Murthy, Ameen, Sanjay & Yasser (2013), identify several licensing models that consumers must consider when deploying applications in the cloud. These include the Pay as You Go (PAYG), subscription, based on the number of users, processor based, based on the number of transactions, based on the subscription to the functionalities, free software with support payments, and the Bring Your Own License (BYOL) licensing models. (Mohan Murthy et al. 2013)

The PAYG model is based on the user's usage. Billing amounts increase alongside the rise of software usage. The PAYG model is specifically useful in scenarios where the number of users is low, and the usage requirement is short term. (Mohan Murthy et al. 2013) Ojala (2013) identifies how the pay-per-use model is a great fit for customers that occasionally need software. In addition, the pay-per-use model prevents vendor lock-in and gives consumers the chance to test and evaluate the software. On the other hand, the pay-per-use model is based on fixed pricing, and it tends to be difficult to predict the usage amount of the software. Another model may be more appropriate for instance if the software is needed on a continuous basis. (Ojala 2013)

The subscription model is typically aimed at users with long term usage of software in mind. As an example, if the user identifies a need of certain software for a predefined number of months, the user must search for the most adequate software subscription choice according to the preferred usage time period. (Mohan Murthy et al. 2013) Ojala (2013) identifies software rental as a subscription fee that consumers pay to use software for a certain time period. The software rental model enables consumers to predict total software costs, as they are contractually defined which prevents the accumulation of hidden costs. However, consumers may end paying regardless of whether the software is used. (Ojala 2013)

The number of users and price increase proportionately for the based on the number of users licensing model (Mohan Murthy et al. 2013). Mohan Murthy et al. (2013) identify this model as cost effective when the number of users is low. Contrarily, the processor-based licensing model price goes hand in hand with processor capacity (Mohan Murthy et al. 2013). Mohan Murthy et al. (2013) identify this model as cost effective when the number of users is high.

In the based on the number of transactions model, the price increases according to the number of transactions being made. On the other hand, the based on the subscription to the functionalities model gives users the flexibility in selecting the preferred modules and functionalities from enterprise software. Therefore, charging is based on the selected modules. In addition, certain software is available for the consumption of the end user at

no cost. (Mohan Murthy et al. 2013) Andrikopoulos et al. (2013) similarly identify how licensing fees are occasionally offered free of charge with accounts. However, support related functions may incur costs (Mohan Murthy et al. 2013).

The BYOL model allows the user to bring an existing license to host an application in the cloud. Another option given to the user is to purchase the license separately, while hosting the application in the cloud. (Mohan Murthy et al. 2013) Similarly, Suleiman et al. (2012) identify how cloud providers offer consumers the option of bringing their own license.

Andrikopoulos et al. (2013) highlight that the costs of software licenses will depend on the provider and the individual licenses of the migrated components. Andrikopoulos et al. (2013) identify that the worst-case scenario licensing wise occurs when migrating the whole software stack of the application to the cloud and not having the ability to reuse licenses. Similarly, SW licensing costs of deployments which include a partial migration of some of the applications functionalities to the cloud are negatively affected (Andrikopoulos et al. 2013). On the contrary, Andrikopoulos et al. (2013) also depict best-case scenarios. Two migration scenarios, replacing components with cloud offerings and cloudifying the application incur the least costs. This is evident as no licenses are required, if the license is included in the pricing model of the provider. (Andrikopoulos et al. 2013)

3. CLOUD CAPACITY MANAGEMENT

Cloud computing has enabled organizations to rent capacity from cloud providers. In some cases, organizations have overestimated their cloud capacity requirements, and matched these requirements based on workload peaks. Having a habit of purchasing more capacity than required can result in major IT budget losses depending on the size of the cloud deployment. Organizations from all walks of industries are estimated to be overspending on cloud services by 42%. (Loten 2018) The cost effectiveness of cloud computing is directly related to the cloud consumers ability to use and optimize the costs of renting cloud resources in a well thought out manner (Sumalatha & Anbarasi 2019). Cloud consumers may formulate management objectives that reflect its approach to resource reservation (Jennings & Stadler 2015).

Hähnle & Johnsen (2015) emphasize the importance of being prepared prior to utilizing cloud services. Hähnle & Johnsen (2015) state that traditional deployment methods are based on specific assumptions, such as the amount of Random-Access Memory (RAM) and CPU. Cloud services on the other hand, offer resource capacity to consumers at a so-called near infinite rate (Hu, Jiang, Liu & Wang 2014). In addition, cloud services give cloud consumers the ability to provision and de-provision resources in a simple and quick manner (Jiang, Perng, Li & Chang 2012). Therefore, alongside cloud computing software needs to be designed for scalability. The designing must be precisely done from the very beginning to avoid accumulating unnecessary costs. The insufficient planning and control of resources is a major reason and barrier to cloud adoption. Tackling these barriers requires awareness on resource consumption. (Hähnle & Johnsen 2015)

The main goal and target of capacity management is to sufficiently maintain optimum and cost-effective resource capacity and ensure that new IT services are not harmed by the inadequate management of capacity (Sabharwal & Wali 2013). Organizations however, exploit varying methods and practices to conduct capacity management and fail to utilize consistent guidelines for managing resources adequately (Lubrecht, Pizzo, Savvides, Baron & Papaefstathiou 2010). Cost savings and optimized resource utilization are said to be the benefits of cloud computing. These correlate with the goals of capacity management. (Sabharwal & Wali 2013)

3.1 Cloud Capacity Management Process

According to Sabharwal & Wali (2013), there are various layers to capacity management that must be considered for overall capacity planning. These layers include business, service and component capacity management. In a cloud environment, not only is the cloud provider held accountable for interpreting business needs and drivers and how these relate to services and infrastructure, the cloud consumer must also have a firm grip on these when deploying resources in a cloud environment. Understanding business activities is the foundation of capacity management in the cloud. These business activities are formulated into organization wide service requirements. In order to meet the necessary service requirements, component requirements must be in line with both service and business needs. (Sabharwal & Wali 2013)

Alongside cloud computing, capacity can be increased on an incremental basis according to business needs. Knowing future business requirements in advance is a prerequisite however, cloud consumption models allow for rapid increases in demand, which therefore reduces the stress of faulty forecasting. This alters the way procurement is handled, as capacity can be adjusted and planned according to demand, without having to forecast capacity needs well in advance like with traditional on-premise environments. (Sabharwal & Wali 2013) Reese (2009) however, emphasizes that when demand is clear, capacity related estimates are more accurate. This assists in avoiding paying for unnecessary capacity or running into the problem of not having enough capacity at hand, resulting in an insufficient infrastructure (Reese 2009). Reese (2009), lists the potential benefits for consumers that have well quantified demand expectations and load estimations:

- Expected loads will be supported in a better manner, as consumers will have better visibility of the resource and infrastructure requirements
- Exceptions are identified when they occur. In other words, when actual load is deviating from expected load.
- Improved understanding on how changes in the application requirements effect resources and infrastructure

Business capacity can have an immense impact on business operations, which is why it should be considered from the beginning of the development process (Allspaw and Kejariwal 2017). The importance of a business aware and holistic method is emphasized in cloud environments, as constant rapid changes take place (Roseline, Tauro & Miranda 2017). From a consumer's perspective, business capacity management in the cloud comprises of focusing on the reduction of the TCO by utilizing the Operating Expense

(OPEX) as opposed to Capital Expenditure (CAPEX) nature of the cloud. In addition, this layer of capacity management considers the potential benefits of the pay-as-you-go model, which works hand in hand with capacity increases that root from business demands. (Sabharwal & Wali 2013) According to Sabharwal & Wali (2013), comparing different cloud service providers to lower costs is also a part of business capacity management.

Furthermore, business capacity management includes capacity planning, which requires consumers to perform business forecasting, financial planning, estimating demand, service level negotiations and application and process re-engineering. These assist consumers in finding a best fit approach for applications deployed in a cloud environment. (Sabharwal & Wali 2013) Amazon further emphasizes how capacity related increases and decreases depend on business requirements and how consumers should only pay for the computing resources that are being utilized by applications (Amazon 2018). Moreover, service capacity management entails ensuring that service-level agreements are being met (Sabharwal & Wali 2013).

Component capacity management focuses on capacity at the component level. Cloud consumers are provided basic component monitoring tools by cloud service providers and third parties to monitor resource utilization and overall performance. These assist the consumer in monitoring component level features, such as CPU and RAM. Tools further enhance and assist decision making to adjust resource utilization at the component level according to the requirements of the previously mentioned service and business capacity. (Sabharwal & Wali 2013)

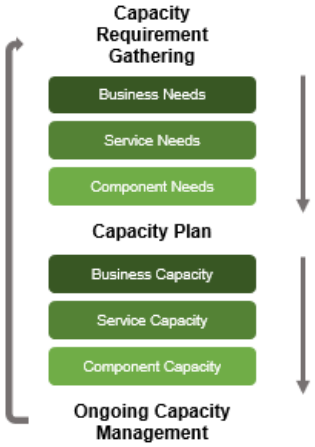


Figure 9. *Capacity plan & ongoing capacity management (adapted from Sabharwal & Wali 2013)*

As demonstrated in figure 9, capacity management starts with the gathering of capacity requirements. These are gathered in order to produce a capacity plan, which then leads

into ongoing capacity management. Although figure 9 depicts capacity management from the cloud providers point of view, the different steps are also apparent for cloud consumers, as discussed in this section prior to figure 9. (Sabharwal & Wali 2013)

3.2 Capacity Management Process

Sabharwal and Wali's (2013) capacity management process prior to moving workloads to a cloud environment is depicted in the figure below.

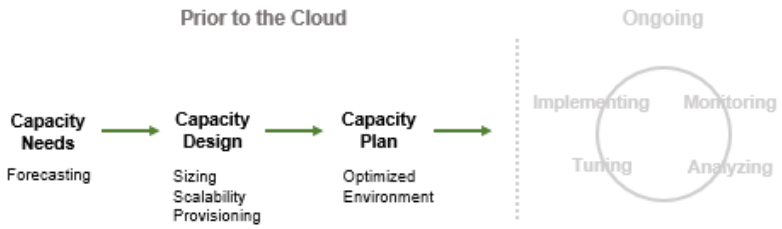


Figure 10. *Capacity management process prior to the cloud (adapted from Sabharwal & Wali 2013)*

Capacity needs should be translated from business to service to component level requirements as discussed in section 3.1. Consumers should have the ability to make appropriate forecasts, as forecasting is essential when attempting to determine capacity requirements. (Sabharwal & Wali 2013) Reese (2009) further emphasizes how consumers must be able to make demand forecasts and understand how they affect applications. Monitoring tools provide consumers with data and information on current utilization and optimization requirements. Therefore, in the case of new cloud deployments, capacity requirements are collected from existing tools. (Reese 2009) Allspaw and Kejariwal (2017) similarly state how capacity planning requires the measurement and historical details of systems and application-level metrics. Microsoft further highlights that in order to estimate costs, understanding of the current resources required to run a workload is important. An inventory of current assets which include servers, Virtual Machines (VM), databases and storage assists in giving transparency to the capacity needs in the cloud. (Micorsoft Azure 2018) Allspaw and Kejariwal (2017) identify a process for capacity forecasts:

- Determining and measuring the essential metric for each resource, i.e. disk consumption
- Identifying constrains of each resource, i.e. total disk space
- Predicting when usage will exceed constrains using trend analysis

Allspaw and Kejariwal (2017) further mention how capacity planning should include the understanding of a systems upper performance boundaries. Knowledge on this matter assists in decreasing the risk of reaching upper boundaries. This includes knowing the fundamental hardware resources, such as CPU, memory, disk and network usage. (Allspaw and Kejariwal 2017) In addition, consumers must try to evaluate future demand in order to forecast capacity utilization. Consumers should have knowledge of how an application experiences seasonal variations or varying demand levels depending on the time of day. It is also important that during this stage of the capacity management process, the consumer weighs the different cost and performance related factors of both on-premise and cloud options. (Sabharwal & Wali 2013)

Singh and Chana (2015) identify that the management of resources in cloud environments is highly complex. The extent of the complexity pushes for the need of efficient techniques to manage resources (Singh & Chana 2015). The capacity design focuses on resource utilization and optimum performance, while keeping costs in mind. The factors included in the design for capacity stage are, establishing a capacity approach, establishing an architecture, applying capacity techniques and checking for cost optimization possibilities. The capacity approach includes being aware of over and under provisioning. Both over and under provisioning of cloud resources effect cloud economies in a negative manner. (Sabharwal & Wali 2013) Faulty provisioning of resources results in wasted time and resources, which cause an increase in costs (Singh & Chana 2015). The capacity architecture should aim at developing applications to consume the lowest amount of capacity, in other words avoiding poor utilization of resources and emphasizing scalability. Applying capacity techniques includes application dependency mapping. Moreover, ensuring cost optimization entails identifying which service and cost models to use according to the applications needs. Cloud consumers have many options to choose from, which further complicates the migration of applications to a cloud environment, as mentioned in prior sections of this thesis. Testing the application in a cloud environment prior to moving the application to the cloud should be considered during the capacity design phase. (Sabharwal & Wali 2013) Similarly, prior to moving any workloads to a cloud environment, the case company identifies how it is extremely important to focus on the design of the application. This makes the biggest difference cost optimization wise, as utilizing cloud capabilities and automation is essential. (Case Company 2019b) In order to take full advantage of cloud computing, consumers must rethink how to design and develop software (Hähnle & Johnsen 2015).

Once the requirements for capacity and capacity design are clear, the capacity plan is formulated as an accumulation of the prior two stages of the capacity management process. The plan should cover the business, service and component aspects of capacity management, and needs to be based on existing and future business demand. The capacity plan should include, but not be limited to, application details, user task scenarios, forecasting, monitoring and metrics. When consumers deploy new applications to a cloud environment, the capacity plan ensures that the maximum benefits are reaped from the use of cloud computing. In other words, the capacity plan aims at ensuring applications are cost efficient. For applications that are already deployed in a cloud environment, the capacity plan aims at keeping capacity requirements optimized and tuned. (Sabharwal & Wali 2013)

Figure 11 demonstrates Gartner's practice-based model of the application optimization process prior to moving workloads to a cloud environment (Anderson 2018).

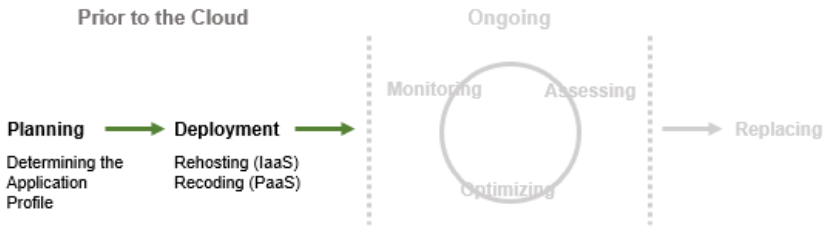


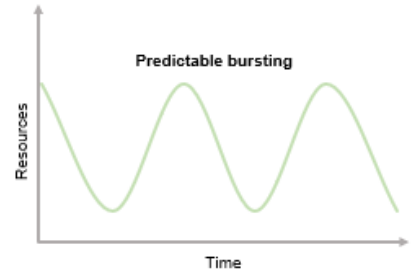
Figure 11. *Application optimization process prior to the cloud (adapted from Anderson 2018)*

The planning phase includes the determination of the application profile. This comprises of sizing the application, mapping all dependencies, identifying all data repositories, identifying integrations and monitoring resource usage. (Anderson 2018) Sabharwal & Wali (2013) similarly agree that when developing new applications, consumers should focus on creating the base case of utilization. These could include factors such as the number of users, or number of transactions. Monitoring tools provide information on current application capacity usage. This data can be used to size the new systems. (Sabharwal & Wali 2013) Both Anderson (2018) and Sabharwal and Wali (2013) emphasize how in addition to gathering all the technical capacity requirements, the planning phase must also take capacity related costs into consideration, while staying in line with business objectives. Wang, Hayat, Ghani & Shaban (2017) further state that services deployed in a cloud environment by cloud consumers must have adequate computing resources to primarily satisfy consumer requirements.

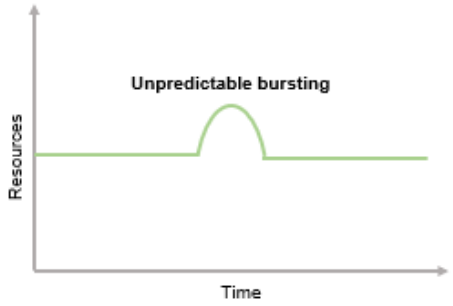
The deployment phase activities depend on the chosen cloud service model. Rehosting (IaaS) means migrating the application. This entails moving the virtualized application

components, re-establishing integrations and moving data or reconnecting application components to data repositories. Recoding (PaaS) considers rebuilding the application. This service model includes the identification of cloud platform services, recoding the application using cloud APIs and relinking using PaaS components. (Anderson 2018) Sabharwal and Wali (2013) also discuss the importance of considering the deployment model during the capacity design phase. Section 2.3 of this thesis covers the different service models in more detail.

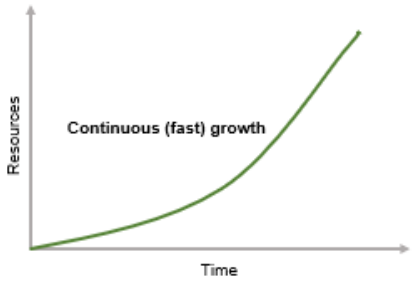
The case company points out various scenarios where moving to a public cloud environment is highly recommended. Figure 12 depicts the different scenarios:



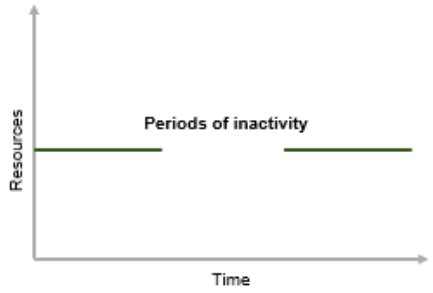
Workloads that experience i.e. weekly, monthly or seasonal variations are most likely best suited for a public cloud environment especially from a service and cost point of view.



The cloud can be configured to handle unpredictable bursts in service demands that originate from varying events.



When future growth of a solution is unknown, the cloud can be configured to scale up alongside the increase in demand. The scaling activity can be automated.



Capacity that is needed on an occasional basis, such as batch workloads that need a lot of capacity at certain points in time can be very cost efficient in the cloud.

Figure 12. *Case company's view on workloads that are the most suitable for the public cloud (adapted from Case Company 2019b)*

3.3 Ongoing Capacity Management

Once applications have been deployed in the cloud, resource utilization data is accessible through various tools. This gives consumers the ability to monitor deployments. Cloud consumers must analyze capacity usage and trends and make necessary capacity related decisions. (Sabharwal & Wali 2013) Post deployment, three continuous activities are performed in order to optimize the costs of the cloud deployment as demonstrated in Gartner's practice-based model (Anderson 2018):

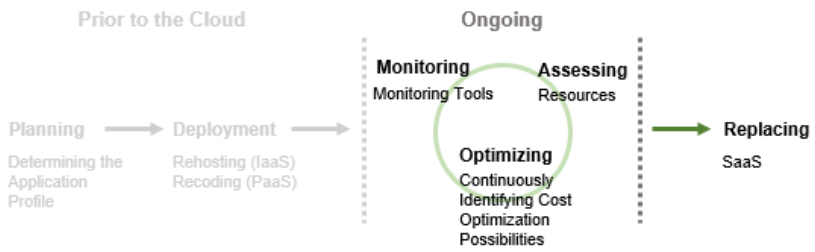


Figure 13. *Ongoing application optimization process (adapted from Anderson 2018)*

Similarly, Sabharwal & Wali (2013) identify continuous activities related to capacity management in the cloud. A total of four activities are acknowledged in figure 14 (Sabharwal & Wali 2013)

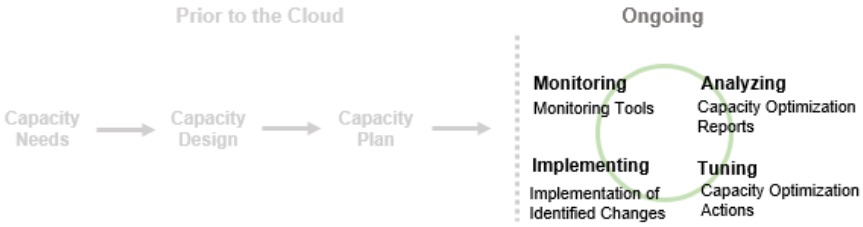


Figure 14. *Ongoing capacity management process (adapted from Sabharwal & Wali 2013)*

Monitoring tools are essential in order to assess resource usage (Anderson 2018). Sabharwal & Wali (2013) agree, that monitoring utilization is vital in order to provide the consumers details on capacity usage. Monitoring enables the consumer to make appropriate capacity related decisions (Sabharwal & Wali 2013). Gartner states how the identification of poorly allocated VMs, resizing and optimizing the VMs and resources as well as continuous assessment of optimization opportunities fall under the ongoing stages of the application optimization process (Anderson 2018). Sabharwal & Wali (2013) similarly identify how analyzing capacity data for forecasting and taking optimization actions to improve resource utilization fall under the continuous optimization activities of cloud deployments.

Replacing in Gartner's practice-based model consists of migrating the entire application to the cloud, in other words using the SaaS service model. Migrating the application is done with the help of migration tools or services. (Anderson 2018) In the SaaS model, data (Case Company 2019b) and users must be managed, as they are the only points of elasticity (Anderson 2018).

Gartner further emphasizes how certain activities should be taken as ongoing actions (Cancila 2015). This includes monitoring and understanding how to spot resources and other items that are no longer being used. This may vary depending on the cloud provider and the available tools. (Blair & Chandrasekaran 2019) Amazon agrees that resources no longer being used accumulate unnecessary costs and need to be identified and removed (Amazon 2018). Furthermore, Gartner states how the identification of unused resources underpinning VMs (Blair & Chandrasekaran 2019) and unassigned storage volumes is important in the cloud. Storage volumes may become disassociated from compute instances. Therefore, monitoring is required for the detection of these storage volumes, as they create line items on the budget. (Cancila 2015)

Scripts and processes should be developed to identify untagged instances as well as idle instances that have not been shut down. These should be monitored on a regular basis, especially the removal of unused instances, as costs continuously accumulate. (Cancila 2015)

Moreover, consumers are given multiple storage tiers. These range from low cost to more performant storage options. Consumers should evaluate the appropriate storage type to optimize storage related costs. (Cancila 2015) Storage itself also requires right-sizing, as there are cost differences between the different types of storage options available (Blair & Chandrasekaran 2019). In addition to storage tiers, consumers should continuously seek new instance types, as they tend to be less expensive and more efficient (Cancila 2015).

Furthermore, the capacity management process should include reporting. This can be handled with the use of appropriate tools. It is important to keep in mind that the correct audience is given visibility to the correct types of data and information during the ongoing optimization activities. (Sabharwal & Wali 2013) Sabharwal and Wali (2013) list the different audiences, and the appropriate reports:

- Business: The business will require budget related reporting.
- IT management: Reporting should highlight and enable the understanding of tactical and strategic possibilities and results that support the business.
- IT operational and technical managers: Reporting should cover component level data in order to provide insight on how to manage resources.

3.4 Resource Management

Resource management can be determined as the process of allocating a variety of resources to a set of applications, in a manner that meets the objectives of several different parties. Resource management has become a challenge alongside cloud computing. Issues root from the scale of modern data centers, the variety of resource types, the unpredictability of load and the altering objectives of the different actors in a cloud environment. (Jennings & Stadler 2015)

From a cloud resource management perspective, Jennings and Stadler (2015) identify three distinct roles that have varying interests regarding resource management. These roles include the cloud provider, cloud user and end user (Jennings & Stadler 2015):

- Cloud Providers: Manage resources in order to provide public cloud services in the form of IaaS, PaaS and SaaS.
- Cloud Users: Ensure that the level of resources leased from public clouds scale according to the applications demand in a cost-efficient manner.
- End users: Generate workloads that are processed using cloud resources. End user behavior can influence resource management, as well as can be influenced by resource management decisions made by the cloud users and cloud providers.

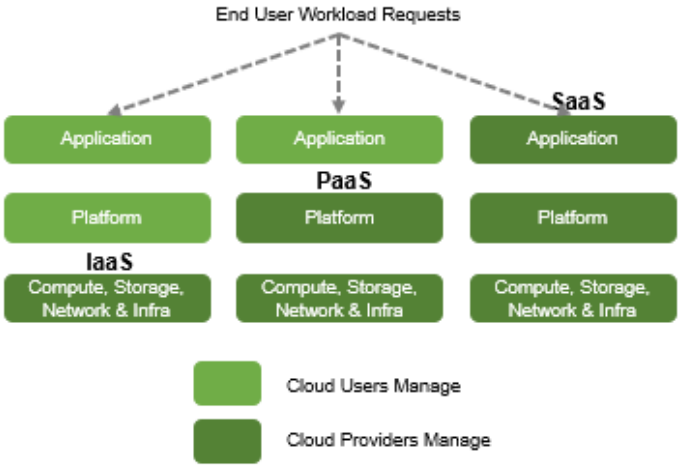


Figure 15. *Resource management (adapted from Jennings & Stadler 2015)*

Resource management is applicable for IaaS, PaaS and SaaS service models however, the level of responsibility resource management wise alters. From the IaaS service model perspective, cloud users are responsible for managing resources according to changing application demands, or for instance changes in cloud provider pricing. (Jennings & Stadler 2015) According to Jennings & Stadler (2015), application scaling and

provisioning fall under the responsibilities of cloud users. Application scaling and provisioning must go hand in hand with demand and change in a dynamic manner. This requires accurate estimations of future demand levels. In addition, cloud users have the flexibility to request or release Virtualized Infrastructure (VI) resources. Cloud users monitor and control the leased resources of the applications deployed in the public cloud environment, as well as control the workloads received from end users. (Jennings & Stadler 2015)

For the PaaS and SaaS service models the functional elements remain the same however, the cloud provider has more responsibility as depicted in figure 15. The cloud user's role in the PaaS service model is split into a platform and application provider. For SaaS service models the platform and application provider can i.e. be the same organization. (Jennings & Stadler 2015)

3.5 Provisioning of Resources

From a business perspective, cloud services deliver utility-based computing (Jennings & Stadler 2015). The cloud assists in avoiding major upfront investments needed for the provisioning of resources (Hähnle & Johnsen 2015), as cloud services enable cloud consumers to have the ability to provision and de-provision resources in a simple and rapid way (Jiang et al. 2012). Therefore, alongside cloud computing consumers have been given the ability to allocate computing resources in an efficient manner, with the luxury of meeting demands (Chaisiri et al. 2009). Jennings & Stadler (2015) state that public cloud computing entails the provisioning of resources to consumers on a leased and usage-basis. In other words, the resources that are required to deliver the needed computing services are measured according to the usage level, duration of use or both (Jennings & Stadler 2015).

Forecasting load is essential in having the ability to efficiently provision resources (Hu et al. 2014). Analytical models can optimize cloud service performance by using accurate predictions. These models enhance the quality of service as well as keep costs at a level that is beneficial for businesses. (Wang et al. 2017) In addition to analytical models, many tools on the market can analyze and monitor resource usage, which in turn assists with forecasting resource requirements (Sabharwal & Wali 2013). Furthermore, Jennings & Stadler (2015) identify how resource allocation predictions are typically based on historical measurements. Similarly, Reese (2009) states how historical patterns form a basis for expectations and forecasts in load, which in turn ease forecasting (Allspaw and Kejariwal 2017). Tools also enable the continuous monitoring of current resource usage

when the applications are already running in a cloud environment (Sabharwal & Wali 2013).

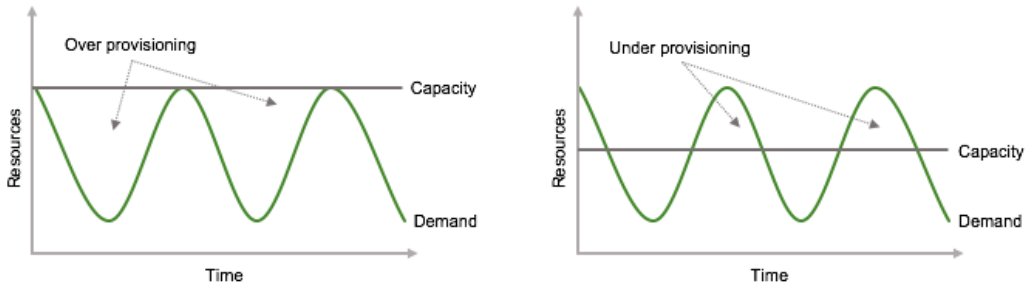


Figure 16. *Over & under provisioning of resources (adapted from Armbrust et al. 2009)*

Over provisioning results in idle resources that are not utilized to the fullest (Hu et al. 2014). Traditional on-premise environments that contain idle resources procured to handle peak time incur constant costs (Sabharwal & Wali 2013). Armbrust et al. (2009) identify how even with the ability to sufficiently provision resources for peak loads, the absence of elasticity results in over provisioning, as demonstrated in figure 16. Majority of the time approximately 10%-50% of a server's capacity is being utilized on-premise (Barroso & Hölzle 2007). Cloud consumers may conservatively over-provision resources to accumulate demand surges, incurring extra costs (Jennings & Stadler 2015). However, Wang et al. (2017) state that having unnecessary VMs results in extra costs. Consumers want to avoid having to pay for resources they do not use in the cloud to combat cost inefficiencies. This will require a change from the traditional way of handling capacity. Consumers must shift from a scaled-up application mentality, to a strategy which opts for the lowest capacity level that supports running application workloads. (Sabharwal & Wali 2013) Another option is to closely match the level of leased resources with demand, which minimizes costs however, risks performance levels (Jennings & Stadler 2015). Hu et al. (2014) similarly identify how under provisioning demonstrated in figure 16 can lead to performance issues, which is another factor that consumers must weigh, when provisioning resources in the cloud.

Resource management requires the appropriate balance between reactivity and proactivity in order to be effective. Reactivity entails the adjustment of resources in response to changes in demand, whereas in the case of proactivity, resources are adjusted in response to predicted demand. (Jennings & Stadler 2015) Proactive instead of reactive provisioning facilitates resource adjustments prior to the time of the load increase. Having the ability to predict application workloads in advance would be the optimal strategy instead of over provisioning as a reactive task. Not only does this cause problems rooting

from the reactive over provisioning, it also results in inefficiencies after the load decreases. The optimal strategy would be to timely adjust resource provisioning according to application demands. (Hu et al. 2014) It is important for consumers to understand the resource requirements of their applications, as cloud services offer different types of cost models and services, which are built to support the capacity requirements of certain types of applications. These include the pricing models, which were presented in section 2.4. (Sabharwal & Wali 2013) As an example, the PAYG model combats the need to forecast load (Amazon 2019).

3.6 Rightsizing Resources

Rightsizing has been defined by Amazon, as the lowest possible resource allocation by cost, that meets the technical specifications of a workload (Amazon 2018). By using the smallest possible amount of capacity feasible for an application, consumers can benefit from cost economies and flexibility. Cloud service providers offer the possibility of very small capacity configurations i.e. in the IaaS service model. This further enables cost efficiencies and ensures that optimum resources are provided for the running of workloads. (Sabharwal & Wali 2013) Rightsizing however, is not a simple task, as modern applications tend to have very fluctuating loads, which result in dynamic resource usage patterns that are difficult to predict and understand over time. It is challenging for cloud consumers to estimate the correct VM size according to the applications load especially when the load is not constant. (Hu et al. 2014)

Sizing environments correctly will be a critical factor in ensuring that the use of cloud pays back businesses as expected (Sabharwal & Wali 2013). Microsoft also emphasizes the importance of rightsizing VMs and storage (Microsoft Azure 2018). Figure 17 demonstrates the possible differences between provisioned, maximum used and average used amount of resources on-premise. CPU/ memory indicates how the differences can occur for both. Falling into the trap of using the provisioned amount will result in a costly cloud IaaS, PaaS or SaaS migration, which contradicts the cost aware business case discussed in prior sections of this thesis. In the example given in figure 17, the PAYG model may be utilized to fill the gap between the provisioned and average used amounts, demonstrating an opportunity to optimize costs. (Blair & Chandrasekaran 2019)

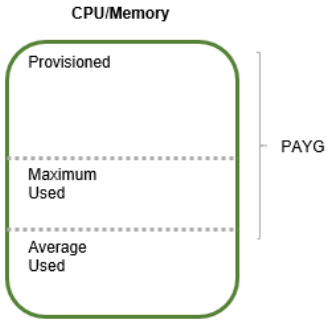


Figure 17. *Example of on-premise resource provisioning (adapted from Blair & Chandrasekaran 2019)*

When in the cloud, rightsizing activity can be triggered by alterations in usage patterns, price drops or new resource types (Amazon 2018). Amazon (2018) lists important areas to consider when conducting rightsizing:

- Ensuring that monitoring covers the complete cycle of a workload over the appropriate time period to avoid faulty provisioning. End user experience must be kept in mind when monitoring.
- Analyzing the cost against the benefit of the rightsizing activity, to assist with prioritization

3.7 Matching Supply and Demand

Matching supply with demand is essential in order to eliminate wasteful provisioning (Amazon 2018). As mentioned in prior sections of this thesis, one of the major benefits of cloud computing is elasticity, as elasticity has the ability to respond to variations in the demand for computational resources. Elasticity gives consumers the ability to avoid the need to excessively over-provision applications resulting in unnecessary costs as depicted in figure 16. In addition, elasticity prevents under-provisioning, which tends to result in a loss of revenue, also depicted in figure 16. In other words, elasticity assists in optimizing resource usage of fluctuating and unknown loads. (Andrikopoulos et al. 2013)

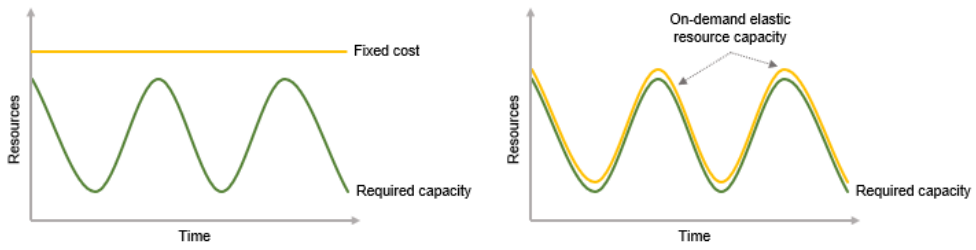


Figure 18. *Economic & flexible resource usage (adapted from Suleiman et al. 2012)*

As depicted in figure 18, elasticity illustrated in the second graph gives the consumer a chance to match expenses with capacity requirements. Scalability is the pre-requisite for elasticity (Vaquero, Rodero-Merino & Buyya 2011). The ability to scale components allows consumers to ensure proper application performance during demand spikes and cost savings during times when demand decreases (Amazon 2018). The fixed cost line in the left graph of figure 18 depicts how traditional on-premise capacity is not utilized to the fullest. Capacity requirements are typically planned according to maximum capacity expectations which requires large upfront capital investments. The elasticity of the cloud allows consumers to avoid over and underutilization of cloud resources. (Suleiman et al. 2012) From a business standpoint, the elasticity of the cloud enables consumers to pay for computing resources only when they are needed. In other words, processing power, memory and additional virtual machines can be added for the client application as needed. (Hähnle & Johnsen 2015)

The cloud empowers consumers to alter computing resources to meet demands with the help of scaling (Reese 2009). Hähnle & Johnsen (2015) identify scalability as another key benefit of the cloud. This entails the automatic adjustment of capacity (Hähnle & Johnsen 2015). Reese (2009) splits scaling into two separate groups, proactive and reactive scaling. Proactive scaling increases capacity based on a plan including projected demand. Reactive scaling on the other hand, increases and decreases capacity by reacting to alterations in demand. (Reese 2009) Scaling can be conducted either manually i.e. through a web interface or executing a command line, automatically i.e. through software that adjusts capacity requirements automatically according to demand, or through predefined changes in capacity. Having the ability to perform manual adjustments to capacity is an improvement in comparison to traditional on-premise capabilities however, the real potential of scaling lies in dynamic scaling. (Reese 2009) According to Reese (2009), scaling in the cloud comprises of three core concerns:

- Having a clear understanding of the expected usage patterns. This may vary according to i.e. daily, weekly or monthly usage, as well as seasonal variance of a business.
- Interpreting how an application responds to load, in order to have an idea of when and what type of additional capacity is required
- Acknowledging the value of systems to the business, to identify when additional capacity provides value. When additional costs are incurred from the scaling of infrastructure, it is important to ensure that the additional costs support objectives.

Gartner identifies how cloud providers offer several services including autoscaling, that allows capacity to grow alongside workload requirements (Cancila 2015). To reap the benefits of autoscaling, applications must be designed accordingly. Autoscaling is an appropriate fit for applications that experience hourly, daily or weekly variability in usage, as demonstrated in figure 18. (Sabharwal & Wali 2013) Having the right amount of resources at the right time can be achieved with the use of autoscaling. Autoscaling monitors applications and adjusts capacity automatically. With autoscaling applications can maintain steady performance while accumulating the lowest possible costs, as only resources that are used are paid for. Capacity alterations are automatically handled in real time according to fluctuations in demand and desired performance levels. Interfaces enable the building of scaling plans and consumers have the option of optimizing performance, costs or both. (Amazon 2019b)

Furthermore, in addition to autoscaling, instance scheduling enables consumers to save costs by up to 70%. This can be achieved by stopping resources that are not in use and restarting them when they are needed again. (Amazon 2019c) Shutting down systems that are idle during the night can enable substantial cost savings (Muhic & Bengtsson 2019). Figure 19 depicts a workload that experiences a halt in demand during a certain point in time. Custom start and stop scheduling can be used for i.e. development and production environments with appropriately tagged instances. Having the ability to shut down instances when they are outside of regular business hours decreases the overall operational costs. (Amazon 2019c)

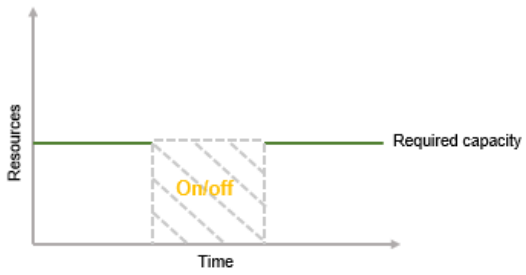


Figure 19. *Example of a workload with no demand at certain point in time (adapted from Anderson 2018)*

Figure 20 depicts an example of how an unexpected spike in demand may occur (Amazon 2018). With appropriate capacity management the expected capacity can be planned for in advance, the unexpected can be recognized in a better manner, and deviations can be reacted to in a more controlled way (Reese 2009).

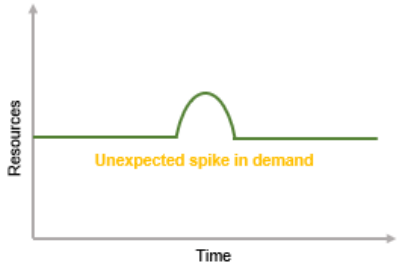


Figure 20. *Example of a workload with an unexpected spike in demand (adapted from Anderson 2018)*

Depending on whether an applications demand is fixed, or variable, automation and appropriate metrics should be adopted to ensure that management of an environment is minimal (Amazon 2018). In addition, elasticity policies should be based on the applications workload changes, as well as business and technical metrics (Suleiman et al. 2012).

4. METHODOLOGY

This chapter describes how the empirical research was conducted. Chapter 4.1 introduces the case company and case study research. Chapter 4.2 introduces the framework used in this thesis, as well as how the framework was used. Chapter 4.3 discusses how the empirical data was collected, followed by chapter 4.4 which describes how the collected data was analyzed.

4.1 Case Study

The case company is a large international industrial company with several different business groups. The case company strives to build solutions which support a sustainable future. The sales of the case company in 2019 totaled to around 10 billion euros and the case company employed approximately 19 000 employees. (Case Company 2019a) The IT department of the case company is centralized and employs approximately 230 individuals and operates in a multivendor environment. Majority of the applications in the different business groups of the case company have been supported by on-premise efforts. However, the shifting of applications to a cloud environment has begun within the case company. The goal of the case company is to move every application to the cloud that can be supported by a cloud model. For this reason, cost optimization and capacity management have become evident. A business process is needed in order to ensure both cost optimization and capacity management are covered within the case company.

Business processes are a topic of interest in case study research with the ultimate purpose of creating new knowledge. Case study research in business and management can accomplish several goals. These depend on factors such as the research questions of the study. (Wiebe, Durepos & Mills 2010) This thesis has a total of two research questions, with the ultimate combined goal of creating a business process. This is one of the main reasons why a case study research format was chosen.

Case study research questions are often focused on business- and management-related phenomena. This however, does not specify that the research questions root from a business development or management point of view. The questions can stem from employees, customers, consumers and society in general. (Wiebe et al. 2010) The research questions in this thesis originate from a variety of perspectives. In the context of this thesis, cost optimization is important from a managerial and business development point of view, making it an incentive for employees of the organization. Capacity management

also includes a variety of viewpoints. Society has identified an issue with the cost of cloud computing. Employees need to have the ability to control cloud related costs with the help of appropriate capacity management. Overall, capacity management becomes a managerial and business development incentive, as it directly correlates with cost optimization.

Case study research is typically impossible to conduct with a quantitative research approach. Instead, case studies support a qualitative research approach with the idea of producing detailed and holistic knowledge, that roots from the analysis of rich empirical data. (Wiebe et al. 2010) Therefore, the aim of this thesis is to create a business process with the combined help of the theoretical background and interviews conducted with employees from the case company.

4.2 Process-Oriented Knowledge Management

This thesis draws from PKM in order to design a business process that takes cost optimization and capacity management factors into consideration in the planning and run phases of an applications cloud journey. The PKM framework was chosen as it designs business processes while keeping value in mind.

Knowledge Management (KM) often lacks strategic perspective within organizations. A variety of theoretical approaches, practical activities, measures and technologies are used for KM, which often results in neglecting to take business and strategic values into consideration. Therefore, more attention should be directed at the strategic value of KM initiatives, as well as the relationship between KM activities and business strategies. (Maier 2002) Furthermore, KM can be classified into two sub-categories, human and technology-oriented KM. To bridge the gap between human and technology-oriented KM, Maier & Remus (2003) suggest a PKM approach. Human- and technology-oriented KM initiatives will still be apparent even when bridging the gap. Both initiatives must extend to include instruments, roles, tasks, contents and systems that are linked and contextualized to enable facilitated navigation in both directions. In addition, an organizational culture that supports closing the gap is also an essential factor in bridging the two initiatives. These can be achieved by redesigning knowledge-intensive business processes and designing knowledge processes which provide an integrating platform for the links and contextualization. (Maier & Remus 2003)

PKM is defined by Maier & Remus (2003) as the management function responsible for the selection, implementation and evaluation of PKM strategies. PKM strives to better organizational performance by improving an organizations way of handling knowledge.

Process Management (PM) and KM initiatives are typically the starting points for the implementation of a PKM strategy. Moreover, the PKM approach is defined by four key levels of intervention, which include strategy, organization and processes, topics/ content, and instruments/ systems. Each level of intervention must work hand in hand to enhance the flow of knowledge within and between business processes. (Maier & Remus 2003) This closely resembles the knowledge lifecycle demonstrated below by Nissen, Kamel & Sengupta (2000).

Figure 21. *Knowledge lifecycle (adapted from Nissen et al. 2000)*

PM initiatives stem from organizational units or process management specific projects. Improvement of process visibility by modeling business processes and knowledge process reengineering are examples of process management initiatives (Allweyer 1999). KM initiatives on the other hand, root from KM projects. The implementation of a Knowledge Management System (KMS) to support business processes covers a technology driven approach to KM initiatives. A KM approach focuses more on comprehensive KM initiatives, which concentrate specifically on other levels of intervention, such as organization and processes, as well as KM instruments. (Maier & Remus 2003) In this thesis, the PM initiatives are more apparent in the implementation of PKM concepts. A new business process must be designed. Although PM is the more dominant initiative, KM is also a part of the PKM initiatives. As previously mentioned, KM initiatives focus on organization and processes, and the implementation of KM instruments. In a process-oriented view, these KM instruments are turned into knowledge processes or take part in the redesigning of knowledge-intensive business processes (Maier & Remus 2003).

Knowledge is believed to be one of the most important strategic resources of an organization. Furthermore, KM must be linked to business strategy, the creation of economic value and competitive advantage. (Maier 2002) The strategy entails defining and implementing an appropriate KM strategy, and acts as a guide for the other levels of intervention. The strategy should have the ability to balance both resource- and market-orientation when designing business and knowledge processes. (Maier & Remus 2003) The

strategy of the case company can be considered resource- and market oriented, as both internal and external factors are apparent. The case company has limited the potential cloud providers to Amazon Web Services (AWS) and Azure. The strategy entails creating a business process for applications that are moving to the cloud. The goal is to build a process that results in cost optimization with the help of effective capacity management. This requires adequate visibility to the appropriate knowledge, and a culture that cultivates cost optimization. The combination of knowledge residing in the minds of employees (implicit) and the knowledge gained from the literature review (explicit) will assist in bridging the gap between technology- and human-oriented KM.

Topics/ content includes the interpretation and construction of process relevant knowledge by gathering knowledge about processes. Knowledge about processes typically stem from process models and knowledge that is created and used within processes. Filtering knowledge from internal and external sources of the organization according to specific business process activities can assist in avoiding information overload. (Maier & Remus 2003) The case company has several ongoing projects where applications are being moved to the cloud. Majority of the current knowledge on processes is embedded in the heads and ways of working of the employees. For this reason, qualitative interviews are used to gain access to the implicit knowledge residing in the heads of the employees. The interview template used for the empirical study is a compilation of themes identified in the literature review section of this thesis. As a result of the interviews, knowledge about the processes require identification and explication. Furthermore, the identification of knowledge created within the process itself requires explication. (Maier & Remus 2003)

Instruments/ systems include KM instruments such as knowledge networks, lessons learned, best practices, process communities etc. In addition, the PKM approach includes instruments such as continuous process improvement and process modeling. Roles, responsibilities, activities and resources must be defined for each instrument and joined into knowledge processes. KMS should have the ability to support PKM. The results of the conducted interviews require identifying instruments, activities and processes. The knowledge lifecycle presented in figure 21 should be considered when designing activities and processes. (Maier & Remus 2003)

Knowledge-intensive business processes are specified as core processes along the value chain, using knowledge to create process outputs. Knowledge processes enable the exchange of knowledge between business units and processes. These include processes that support the collection, organization, storing and distribution of knowledge,

as well as processes that manage the allocation of skills and expertise to business processes or projects. Knowledge management processes manage the organizational knowledge base. Focus is placed on the continuous improvement of the knowledge base. KM organization and processes in PKM utilize knowledge lifecycle activities and combine them into knowledge processes. Business processes must also be linked in order to integrate processes and KM. As an example, process manager and knowledge manager roles can be assigned to one person. Furthermore, another example is the enhancement of existing activities within business processes with KM activities. The idea is to bridge the gap between human- and technology-oriented factors into one single process. (Maier & Remus 2003)

4.3 Data Collection

Interviews can be used to collect data as well as gain knowledge and understanding through conversation. Interviews are typically conducted face to face or via different tools used for communication. They are often recorded and further transcribed, creating a data source for analysis. (Wiebe et al. 2010)

Structured interviews were conducted in order to collect the data used for analysis in this thesis. Structured interviews entail giving all interview participants the same set of questions. (Wiebe et al. 2010) The interview questions were based on themes gathered from the literature review. Two separate interview templates were created in order to ensure participants in different phases of the cloud journey were taken into consideration. The first set of interview questions were for individuals in the planning phase of the cloud journey. The second set of questions were for individuals that already have applications in a public cloud environment or are currently in the migration phase.

Case study research often entails purposeful sampling. This form of sampling includes the selection of information-rich cases in order to enhance the understanding of pivotal topics investigated in the study. (Wiebe et al. 2010) A total of 14 interviews were conducted in this thesis. Interview participants were selected by the case company's Cloud Architect. Selected interviewees were at different phases of the cloud journey, either planning, migrating or in cloud, as demonstrated in table 1. In addition, all interviewees are employees in the case company's IT department. Table 1 specifies the role of each interviewee in more detail. A total of three interviews were conducted face to face, and 11 interviews were held via Microsoft Teams, with an average duration of 41 minutes. Each interview was recorded in order to ensure all the data could later be transcribed and further analyzed. Table 1 depicts the interviews conducted in this thesis.

Table 1. *Summary of the conducted interviews*

Interview	Role in Case Company	Cloud Journey Phase	Duration (mins)
I1	IT Architect	Planning	37
I2	IT Architect	Migrating	34
I3	Chief Architect	Migrating	47
I4	IT Architect	Migrating	45
I5	IT Architect	In Cloud	32
I6	IT Service Owner	Planning	38
I7	Senior IT Architect	Migrating	55
I8	IT Service Owner	In Cloud	50
I9	IT Service Owner	Migrating	21
I10	IT Architect	Planning	44
I11	Manager, IT & Digitalization	In Cloud	36
I12	Senior Manager, IT	In Cloud	34
I13	Project Manager, IT	In Cloud	57
I14	Senior Manager, IT	In Cloud	49

4.4 Data Analysis

For the data analysis part of this thesis, the interview recordings were initially transcribed word for word. Once the transcribing was done, the data was grouped into several themes using Microsoft Excel. The themes were based on the themes in the interview templates, while keeping research question one in mind. The data was further analyzed into more specific subcategories beneath each theme. During the categorization, data was combined from different interview questions to form the subcategories.

In addition to the literature review, research question two was also used to formulate several questions for the interview templates. For this reason, a few questions were asked to specifically identify process related implicit knowledge, as well as roles and responsibilities in different situations. Therefore, certain subcategories within the themes were grouped to support the use of the PKM framework.

Once the data had been analyzed, the PKM framework was used to build the business process. Initially, processes and sub-processes were identified. The literature review and empirical results were further combined to detect activities along the processes and sub-processes (business process). In addition, key internal and external resources of the different activities along the business process were identified. Furthermore, a KM process including instruments, tools and knowledge was added to the business process and assigned roles and responsibilities. The findings were gathered using Microsoft Excel. The Excel was then used to draw the processes presented in chapter 6.

5. EMPIRICAL RESULTS

This chapter presents the empirical results that were gathered in the form of interviews. Chapter 5.1 includes the motivation and business justifications related to the different cloud journeys. Chapter 5.2 discusses different topics prior to the cloud. Furthermore, chapter 5.3 depicts relevant topics related to activities in the cloud, and chapter 5.4 discusses a variety of topics related to cost optimization.

5.1 Motivation and Business Justification

The current strategy in the case company's IT department is to move applications from the on-premise data center to a cloud environment. The strategy itself has pushed case company employees to consider cloud solutions. IT Architect (I1) and IT Service Owner (I6) stated how the main reasons for beginning the cloud journey were because of higher IT management and the overall strategy. Although the strategy has played its part in the decision to move to the cloud, it was evident that many other reasons also had a great effect on the decision to begin the cloud journey. Senior Manager, IT (I14) mentioned ease of use, flexible sourcing models, cost flexibility and scalability as business justifications. IT Architect (I2) similarly identified various reasons:

> *The ability to replace CAPEX with OPEX. Scalability, and of course process wise it is simpler… The cost related aspects are a leading factor for us, and there are technical benefits as well. – IT Architect (I2)*

The advantages of scalability were a common theme among interviewees. IT Service Owners (I6) and (I8), IT Architect (I2), Manager, IT & Digitalization (I11), Project Manager, IT (I13) and Senior Manager, IT (I14) all mentioned scalability as one of the reasons to move to the cloud. Especially when comparing on-premise and cloud options, the scalable and flexible nature of the cloud increases its attractiveness. Scalability has also been strongly evident when considering the applications characteristics and the potential benefits scalability brings cost wise:

> *For the third archive model, the aim is to get cost savings as the archive system is rarely used, so it would be built in a way that when the user wants to use it, the user starts up the servers and services in general, and in that way, it is a cost saving. – IT Architect (I1)*

> *From a cost perspective we saw scalability as an opportunity to flexibly scale and turn things off. We could identify this as a good fit for our design software. – Project Manager, IT (I13)*

Manager, IT & Digitalization (I11) mentioned how scalability is a central factor, as the size that their business will grow to in the near future is still unknown. IT Service Owner (I8) further identifies how scalability enables the ease of expanding resources on a need basis:

> *The cloud is more flexible than the on-premise environment. We started off with very minimal resources on our servers to keep our costs low… If we need to add capacity, the additions are only a few simple clicks away. – IT Service Owner (I8)*

The cost of the cloud in general and the cost savings enabled by the technological capabilities of the cloud were mentioned in many of the interviews. IT Service Owner (I6) specifically stated that money and costs are drivers and constantly kept in mind. This however, was often identified as important, but not the main motivator. When asked about the importance of costs, IT Architect (I4) stated:

> *It is, but it is not the motivation. The motivation is agility. Well technology in general. So, I mean it is beginning to be more difficult to find vendors that do not work in clouds. IT is developing, this is the new modern way of doing things. – IT Architect (I4)*

Keeping up with technological advancements such as the cloud has been acknowledged as something that the IT department needs to primarily take care of. When specifically asked for business justifications, Senior Manager, IT (I14) mentioned how in principle it is not that relevant for the business whether the solution is on-premise or in the cloud. In other words, the technical journey may not be the main topic of interest for businesses, and therefore IT needs to consider the technical side of things. Oftentimes the targeted business outcomes are reached with modern technologies such as the cloud, as stated by Senior IT Architect (I7):

> *This project was primarily coming from IT, but of course even IT projects need to have some business justification. For the business the most important reason to go there was the performance…Then the other thing was the simplification of the landscape… We were also kind of promising business that the more you simplify the landscape the more robust it is. – Senior IT Architect (I7)*

Similarity to I7, IT Service Owner (I9) highlighted performance and simplification of the whole architecture as the major expectations from the business and advantages of the

cloud. I9 also added, that in addition to business justifications, IT itself may run into a situation where IT justifications need to be analyzed:

> *IT side justification is that we had an ageing platform. The whole platform is close to 6 years old and nearing end of life, so we had to decide whether we renew the same on-premise type of platform again, or should we move with the future in mind to prepare ourselves for the cloud journey. – IT Service Owner (I9)*

Nearing end of life was mentioned by a few interviewees as a turning point to shift applications from on-premise environments to the cloud and modernize the way things are done. Modernizing the current environment was another major theme that was evident from the interviews. For IT Architect (I10) for instance, renewing the application and rebuilding its functionalities enabled the use of the application on a global scale. Globalizing the application was a business case that could not be reached with the old software. Moreover, Chief Architect (I3) discussed how there has been a vision to do things in a more modern way for a while, and how IT should be an enabler of agility. This has therefore led to the increasing use of the cloud. IT Architect (I5) mentioned how moving to a cloud environment was a solution to simplify the deployment of mobile applications. I5 emphasized how a cloud environment facilitates the development and deployment of applications for vendors and developers, and how the simplification brings cost efficiencies.

Furthermore, Senior Manager, IT (I12) discussed how the shift towards a cloud environment is an investment itself, that different areas of business can capitalize on:

> *IT justified the investment of the platform by stating that it increases the ability to use data and fosters the maturity of data. The benefits are not direct, instead they are indirect in a sense that we can offer businesses much better data skills and data technologies, which in turn enhances their data skills, and that way the business benefits are identified from the actual use cases. – Senior Manager, IT (I12)*

5.2 Prior to the Cloud

During the interviews the interviewees were asked to talk about the activities they are currently conducting or conducted prior to moving applications to a cloud environment. The questions mainly dealt with forecasting and estimating capacity, the tools used to estimate capacity and spend, as well as any problems that occurred with the capacity estimations. Table 2 summarizes the following sub-sections by listing the topic in question, a short description of the content and either a direct citation or short description of the results.

Table 2. *Summary of the prior to the cloud interview results*

Topic	Description	Results
Estimations	Ways to estimate capacity needs	*Understanding of the cloud possibilities and your target design is really something which will help you make the capacity estimations in a proper way - (I7)*
Tools	Tools used to estimate spend & capacity	*Excel to calculate it, but the source of course is the AWS price list - (I7)*
Problems	Issues with capacity management forecasts and estimations	*It was proven that when you know how to make a smart design, it will significantly decrease the amount of capacity needed than if it were done in a way that first comes to mind - (I11)*

5.2.1 Estimating Capacity Needs

Making capacity estimations prior to moving applications to the cloud varied depending on the type of deployment and the nature of the application. The chosen service model for the cloud journey also required certain different methods to estimate capacity. In addition, estimating capacity slightly altered depending on whether there was a similar solution on-premise prior to moving the application to the cloud.

Majority of the interviewees who decided to use the IaaS service model were fairly confident about the ability to estimate capacity needs. The current state of the cloud journey, either planning, migrating or already in a cloud environment did not seem to affect the ability to estimate the capacity needs of IaaS deployments. This rooted from the fact that many of the IaaS solutions were going to be moved or had already been moved from the on-premise environment to the cloud. IT Architect (I1) stated how there are no issues with capacity related estimates, however If something new was being built it would be more challenging. Similarly, IT Architect (I2) mentioned that capacity estimates can be taken from historical load estimations of information systems. I1 and IT Service Owner (I6) further depicted:

> *When it comes to legacy systems it is pretty clear, accurate figures exist. – IT Architect (I1)*

> *For this first case we have easy access to the requirements on how many and what types of servers, disk space and other things we need. This is based on the current production environment. – IT Service Owner (I6)*

Some interviewees received the capacity estimates from relevant vendors. For IT Service Owner (I8), the vendor created estimates of the capacity requirements. These were also based on prior experience of the applications resource usage. Similarly, Project Manager, IT (I13) received the capacity needs from a vendor, which were based on a sizing exercise that determined the appropriate amount of CPU, RAM and other resources. I8 further described how the capacity was estimated:

> *It was based on the number of simultaneous users and the document flow. Approximately 600 000 pages are processed a year, which accumulates to around 2.5, 3 hundred thousand separate documents. – IT Service Owner (I8)*

Furthermore, understanding the capacity requirements to begin with are important however, it is essential to recheck if the capacity needs are still similar to those on-premise. IT Service Owner (I6) identified how for one application, load and usage levels will be much lower in the cloud, than in the current on-premise production environment. IT Service Owner (I9) further recognized how capacity needs should be thoroughly analyzed, as this may lead to reductions in capacity requirements. Senior IT Architect (I7) emphasized the importance of design:

> *When we were in the planning phase, that time we had the understanding that for production we need 4 TB, and for the rest of the six test environments we needed*

> *2 TB. However, during the project phase we also did some more analytical understanding and checks, and then it came out to be that right now we are in the production with 2 TB and all the test systems are 1 TB. – IT Service Owner (I9)*

> *This capacity is directly related to the design. In fact, first you do the design and based on this design you see what capacity you need… Understanding of the cloud possibilities and your target design is really something which will help you make the capacity estimations in a proper way. – Senior IT Architect (I7)*

Although capacity estimates were a common activity prior to moving applications to an IaaS cloud, Project Manager, IT (I13) reminded that the cloud itself is designed to ease the provisioning of servers. There is no major wait like on-premise, and no need to worry if certain components were missed in the delivery. He further stated how this minimized the risk of wrongly estimating capacity requirements:

> *You can always crank it up or down… You can just test it and if it is not suitable, then add some more. – Project Manager, IT (I13)*

IT Architect (I4) and Manager, IT & Digitalization (I11) similarly identified how the cloud enables the agile use of resources:

> *Well when we talk about IaaS then basically that is where the cloud gives us agility, so instead of ordering a server and kind of predefining it, you could actually measure it as you go, and it is easy to change it and adjust when you need. Because when we talk about servers, so of course there is some reference of the physical hardware we had beforehand and that gives general guidelines. – IT Architect (I4)*

> *We started out with a minimal setup which is increased on a need basis. The smallest standard servers have been taken into use…Then if problems come up, we add more HW. – Manager, IT & Digitalization (I11)*

For PaaS deployments the capacity management side of things was still very new to some interviewees. IT Architect (I4) discussed how they are taking the process step by step, as they are still learning what it means to use PaaS in the cloud:

> *We need to learn it, so we take small cases. We kind of design it and we learn as we go, what does it mean for capacity management… The mental shift in change. Its capacity management from all sorts of aspects, so it is the financial capacity like how much you end up paying for it, but it is also the technical limitations as well as build pipelines and processes etc. The journey will not end this year. – IT Architect (I4)*

Starting out with minimal resources and the cheapest plans was a common theme for PaaS deployments. Chief Architect (I3) mentioned how it may be difficult to predict the future, but a clear roadmap should be made to identify if additional resources are required down the line. Manager, IT & Digitalization (I11) pointed out how there are price lists that show how much different components cost, and by starting from the lower end of the scale, the costs are instantly known. IT Architect (I10) pointed out that instead of servers, native components were taken into use. The components have tiers which can accumulate major expenses if used incorrectly. I10 knew the number of users and how active the users were with using the application. With this knowledge I10 created a technical solution to understand the capacity requirements:

> *We created DEV, QA and PROD environments for our application. We started out by making the DEV side as cheap as possible and observed how it works and if the capacity was enough for running it. Then we moved it to the QA environment to run it, and then for the PROD environment we cranked the capacity up a couple of notches… With this we were able to see how much capacity we need etc. – IT Architect (I10)*

Whether or not the exact capacity requirements are known prior to shifting applications to a cloud environment, certain features of the cloud can assist in gradually increasing capacity according to needs. This is especially apparent when minimal resources are taken into use in the beginning. IT Architects (I5) and (I2) both mentioned how autoscaling can be set to assist with the management of capacity:

> *What we do to focus on cost optimization is that we always start with the cheapest plan… We always start with the more standard plan. We set up the autoscaling and if we see that something is scaling too much or something like that, then we scale one step up. But most of our plans are standard plans. – IT Architect (I5)*

> *I can throw an estimate which helps us to get started. Worst case scenario if estimating goes wrong, we can still react rapidly, as we are not tied to any physical solutions, so instead we can drive them up or down in the cloud. There should always be some estimate on the number of users… Also, during the development phase, you should have some understanding on how it uses resources. As an example, in Azure you can set limits to avoid having to scale, and instead the service scales itself between a given minimum and maximum amount. – IT Architect (I2)*

Chief Architect (I3) mentioned how for new services it is difficult to know the workload on the resources. In these types of cases I3 has started by estimating what the workload

would most likely be. In addition, I3 continued that services with unpredictable spikes and moments of very little usage are hard to estimate. Furthermore, I3 talked about a serverless DB, Azures Cosmos DB. I3 stated that once a DB account is made the user can set the amount of resources (resource units) they want, and then if 100 are selected, the user can make a certain amount of calls within a minute to that specific database API. I3 emphasized how in the future a slider could be programmatically set according to the workload however, for now I3 has been utilizing application service environments where a certain amount of capacity can be bought and reserved. The load is rather steady for I3, so this system has worked until now.

IT Service Owner (I6) and IT Architect (I10) further clarified the need to understand the user amounts:

> *We had parameters on how many users we have and how we want to use it. – IT Architect (I10)*

> *The user amount is one thing which most likely also affects the capacity estimates. So, clarifying the user amount for the capacity is probably another thing. – IT Service Owner (I6)*

5.2.2 Tools

The tools used for calculating spend and capacity related forecasts were the same for each service model. The cloud providers online documentations and calculators were used to identify prices. Microsoft Excel was used to make the relevant calculations. Senior IT Architect (I7) summed up the activity that majority of the interviewees conducted:

> *Excel to calculate it, but the source of course is the AWS price list. – Senior IT Architect (I7)*

IT Architect (I4) gave examples on the type of details that can be found online:

> *Well the prices are known I mean if you use a Lambda function you know how much it costs for every second it runs and, in that sense, even Amazon and Azure they offer these calculators… As well as the documentations for technical limitations, no more than X Lambdas per region running at the same time and those types of things. So basically, the cloud documentations, I guess calculators are a part of that in a way, but no tool as a kind of external tool to do something smarter. – IT Architect (I4)*

According to Senior Manager, IT (I14), forecasts for SaaS solutions are straightforward:

> *We used Excel… to compare the prices, as in how much it will cost if there are this and these many users. But it was purely SaaS, so it was straightforward. – Senior Manager, IT (I14)*

Forecasts that were created by vendors were also based on the cloud providers documentations. IT Architect (I10) mentioned how one vendor used the Azure online calculator to estimate the costs of the components that were needed. On the other hand, according to Senior IT Architect (I7) vendors occasionally have their own sizing tools as well.

Project Manager, IT (I13) mentioned how the same tools were used for the forecasts as by majority of the other interviewees, but certain surprises came up along the way. The case company had different pricing than what was available online. However, when the service price of the case company's cloud service partner was added, it approximately made the prices the same as online. In addition, there was no knowledge on the fact that reserved instances could decrease the price, so this was not taken into consideration in the initial calculations.

Moreover, certain interviewees in the planning phase were unaware of the available tools provided by the cloud providers. This demonstrates how there may be a lack of communication or process knowledge regarding the cloud journey, as these tools were not evident to begin with. IT Service Owner (I6) was unaware of any tools and mentioned how it is hard to get started especially when the topic is new:

> *No understanding of tools, and I don't know what type of tools they usually are…This week I saw where server costs could be found. It was some Azure calculator. – IT Service Owner (I6)*

5.2.3 Problems with Capacity Management Prior to the Cloud

There were certain problems related to capacity management that were identified by the interviewees. Forecasting was at times difficult for a few different reasons. Two of the interviewees, IT Architect (I4) and Senior Manager, IT (I12), offer services to different business areas within the case company. In other words, they offer shared services to a large audience. Therefore, it is rather challenging to predict usage levels beforehand, as there is no proper visibility of the needs. I4 however, identifies this as a problem that has been present in the on-premise environment as well:

> *Mainly that we don't know the forecast but that's true not only in the cloud. Just the fact that we are a shared service providing services to all business towers, but we don't know in advance their needs. – IT Architect (I4)*

Senior Manager, IT (I12) similarly struggled with knowing and predicting the usage levels. I12 had no idea of the type of use cases and the number of users that would be using them, making it extremely difficult to estimate the necessary resources, such as storage capacity and computing power. I12 further described how many of the services used are pay as you go, and for that reason a decision was made to always give a disclaimer:

> *We always had to give the same disclaimer, that the cost will purely depend on the actual use, and for this reason we could not give a direct budget, that this is the amount of money we are going to use on the infra. Instead it was always done with a disclaimer, that the estimate is this, and the actual cost will depend on the use. – Senior Manager, IT (I12)*

According to Senior Manager, IT (I14), forecasting can go wrong in two different ways. The capacity needs could be wrongly estimated from a technical standpoint, or from a timetable point of view. I14 ran into an estimation issue, which negatively affected the budget. The capacity needs had been planned and budgeted however, the estimations were made without any buffer. Therefore, the budget exceeded as the servers needed to be up and running for a longer period:

> *If we have planned a certain capacity need, or that our servers are up and running for instance from 8 am until 6 pm… if there is a need to work on them longer and the servers are up 2 hours longer, for let's say 2 months… the budget will quickly exceed, if the servers are needed to be kept up and running for longer. – Senior Manager, IT (I14)*

According to Manager, IT & Digitalization (I11), a very spikey load can make it harder to make capacity estimates. I11 talked about how the usage of a reporting service is at times very heavy especially when data is being downloaded, so during those times the capacity is at its limit. Then during other points in time there is no usage of the service at all. If someone were to use the services in between the activity then there may be extra load, making it hard to estimate and optimize. I11 further emphasized how even the slightest changes may yield great improvements in performance. These however, may require someone who is very experienced and familiar with the case at hand:

> *It was proven that when you know how to make a smart design, it will significantly decrease the amount of capacity needed than if it were done in a way that first comes to mind. – Manager, IT & Digitalization (I11)*

IT Service Owner (I9) identified certain issues related to changes in technology. According to I9 it was difficult to understand and know the amount of decrease in capacity, when shifting from one solution to another:

> *The major problem has been when we are moving from Oracle to Hana database, and Hana is actually needing less capacity than in the Oracle world, and I think the problem has been to understand that what is that less. – IT Service Owner (I9)*

Moreover, in one of the cloud journeys, an accident occurred where autoscaling was thought to be in place, when in fact it was not. Corrective actions were taken to ensure that autoscaling was working on all the applications to avoid the incident from taking place again:

> *We did have some issues with this one time because the vendor… did not set the autoscaling as it should have been set and we had some services shutting down because too much traffic, and it was not able to handle the workload. – IT Architect (I5)*

5.3 In the Cloud

During the interviews the interviewees were also asked to talk about the activities they conduct when the application is in the cloud. Thoughts and opinions were gathered from interviewees at the very beginning of the cloud journey, whereas actual current ways of working were gathered from interviewees with applications already in the cloud. The discussion mainly included the frequency of capacity management activities, monitoring and tools, visibility, current issues, exit plans and roles and responsibilities. Table 3 summarizes the following sub-sections by listing the topic in question, a short description of the content and either a direct citation or short description of the results.

Table 3. *Summary of the in the cloud interview results*

Topic	Description	Results
Frequency	How often capacity management activities take place	Costs on a regular basis, performance based on need
Monitoring & Tools	Tools used to monitor capacity, usage and spend	AWS and Azure native tools mainly used
Visibility	Visibility of cloud resources (spend)	*We have no visibility to the costs - (I10)*
Problems	Issues in the cloud	Tags, growing usage and capacity amount, understanding cloud bills
Exit Plan	Methods and opinions on exit plans	*You should think of an exit plan regardless of whether the costs are rising or not - (I4)*
Roles & Responsibilities	Capacity plan/ design	Case company, vendor, or both
	Monitoring	Case company, vendor, cloud service partner
	Implement change	Vendor, cloud service partner

5.3.1 The Frequency of Capacity Management Activities

Capacity management related activities in the cloud did not really have any specific uniform pattern on how often the activities take place. This partially rooted from the fact that some of the interviewees were still very new to the cloud environment or still migrating, and because the applications of the interviewees varied. Service models also affected the capacity management routines in the cloud.

Although no clear routine was in place for several interviewees, the individuals were still aware that there is room for improvement when it comes to the capacity management

activities in the cloud. Senior Manager, IT (I14) stated how it would be important to do this on a regular basis, to see whether the capacity is adequately used. IT Architect (I5) mentioned how the activities do not take place that often however, I5 identified that it is something that should happen more frequently, so improvements are needed. IT Service Owner (I8) similarly stated that this is something they would like to do more. IT Architect (I4) further mentioned how a lack of time negatively effects this:

> *So exactly we don't have a vendor we don't have time for it, that's the more concrete answer. We would like to do more. So today it's done mainly when we hit a wall, like whenever there are issues. But as said its static, so unless something has changed in the environment, suddenly more messages or something, then it's kind of left statically. – IT Architect (I4)*

IT Service Owner (I6) did not take any stand on the regularity of the activity, as at this point in time the cloud journey is still in the very beginning. I6 believed that the activity could depend on whether capacity was being added or decreased:

> *For instance, if something needs to be expanded, these types of things need to be monitored constantly to avoid having a situation where we run out of space somewhere, and that causes the service to turn off. But then if we think about this from another angle and there is a need to decrease something, then that could take place less frequently. – IT Service Owner (I6)*

For several interviewees, capacity management activities took place on a need basis. For Manager, IT & Digitalization (I11), this on average happens once or twice a year, but there is no regularity to it. According to I11, typically this activity originates from a need, when for instance end users alert that something is not working, or the interviewee catches the issue before the end user. I11 has a dashboard with all the relevant information, so changes in the behavior of the application can be identified. Project Manager, IT (I13) identified this taking place on a monthly basis or once every two months during the build phase rooting from factors such as slowness. This however, is believed to decrease once the application stabilizes. For I11 and I13 the frequency of activities varies. Issues related to performance are based on need however, costs are checked on a monthly basis. Chief Architect (I3) stated that on average the capacity and budget are checked once every three months. Moreover, Senior IT Architect (I7) split costs and performance into two separate categories:

> *I would say that at least monthly we can check it, well monthly we do it for sure, I am checking the costs. So, from this perspective monthly, but when it comes to*

> *the performance then based on the need. We noticed during our UATs for instance that in some cases operations were not possible because the storage was not sufficient, so we had to extend the storage. We added one application server but that was kind of based on the need, like it was constant monitoring in a way. – Senior IT Architect (I7)*

On the other hand, certain interviewees believe that the design of the application plays a role in the frequency of capacity management activities in the cloud:

> *Hard to say, hopefully never as in the application would be designed so well that it is not necessary. A yearly check would be good. – IT Architect (I1)*

> *Very Rarely. I cannot recall having to go and do anything manually… we don't want virtual servers and we don't create any, so then we don't have to adjust them. When cloud native things are done, it just works. – IT Architect (I2)*

The nature of the application also determined the outlook on how often capacity management should take place. According to IT Service Owner (I9) it is still difficult to say, as the application has not yet gone live however, the test systems would need to be in use at least three times a year for two or three weeks. Therefore, the activity would take place accordingly, but mainly on a need basis. Senior Manager, IT (I14) suggested that the regularity of the activity could depend on the capacity pricing. For instance, if the capacity follows a minute-based pricing system then a monthly check could be too infrequent. Then again if the pricing was daily or monthly, it would require something different. Furthermore, the size of the application may also affect the regularity. IT Architect (I10) mentioned that the application is so small that there has not really been a need so far. On the other hand, for Senior Manager, IT (I12) capacity related changes are constantly taking place, as new things are continuously being developed. For this reason, the activity takes place on a weekly basis, according to need. I12 was not able to clearly identify the cycle of this activity in the future, but most likely it would be similar to the current situation:

> *Development takes place, user amounts change, there is more data, so there is a need to do something the whole time. Small constant changes. – Senior Manager, IT (I12)*

5.3.2 Monitoring and Tools

The case company has two main cloud providers, Azure and AWS. For this reason, majority of the interviewees rely on the native AWS and Azure tools to monitor capacity, usage and spend. Manager, IT & Digitalization (I11) discussed how the Azure portal

clearly shows per subscription what there is and how much it costs. In addition, IT Architect (I2) mentioned how Azure shows the costs and spends according to the components. Chief Architect (I3) similarly stated that the API management dashboard is used to follow key metrics on the utilization of capacity. Additionally, I3 follows the Azure budget on a monthly basis to compare the forecast to the current situation.

A couple of interviewees mentioned application specific self-monitoring and technology specific tools that are used in addition to the cloud native tools. Senior IT Architect (I7) summed up how different tools are used for system performance monitoring, usage monitoring as well as spend monitoring in their area:

> *SAP native tools to monitor the system performance… AWS tools to also monitor the usage. For costs there is this cloud service partner dashboard, which is fetching the costs from our accounts in the case company, so this is the only tool, and then there are the AWS tools… Right now, the target is to use this cloud service partner Insight. – Senior IT Architect (I7)*

One thing which was evident from many of the interviews was that monitoring seems to be at very different levels depending on the phase of the cloud journey. Some interviewees have placed alerts for resources such as CPU and memory to notify when the set thresholds are exceeded. Project Manager, IT (I13) set up the alerts early on. On the other hand, some were not aware of how the monitoring is being done but knew that the agreed upon standard tools were used by the vendor. Moreover, for Senior Manager, IT (I12) monitoring is currently being worked on with a vendor. In I12s area, everything is monitored separately, and for the time being there is no centralized monitoring.

IT Architect (I5) on the other hand had set up a different type of monitoring system:

> *We have Azure monitors send emails when the capacity is triggering some alerts, then basically the stakeholder that needs to receive this email is notified… Usually what we are monitoring is the resource usage and if they are going above of what we expect, and if something is coming down. We have some issues sometimes with the storage accounts and also if they are shutting down, we also have some alerts to notify that. – IT Architect (I5)*

The monitoring also varied depending on the types of cloud deployments. IT Architect (I4) emphasized how for EC2 servers (IaaS), it is evident that usage of CPU and memory are followed and the tool that is being used has internal reports and self-health checks. On the other hand, IT Architect (I2) is using Splunk to follow logs etc., and emphasized how not having any virtual servers entails not having to check on i.e. CPU usage:

Cloud native functions do not analyze who does a call, instead 1 call is equal to 1 cent. – IT Architect (I2)

5.3.3 Visibility

Visibility was a topic that was brought up by several interviewees. For IT Architect (I1), the on-premise visibility is not that great, and the current system usage amount is unclear. I1 emphasized that disk space can be seen, but utilization amounts, such as the number of users that have been there, and for how long they have used the system are lacking visibility. I1 wonders if the cloud will change this:

Not sure if the cloud changes this. It would be good that unnecessary capacity which is not needed can be removed, or then increased if during some points in time there is heavy load. Or if on the weekends it is not used at all, some timings could be checked when services are turned off. This affects the costs. – IT Architect (I1)

According to Project Manager, IT (I13), there is currently no visibility of the cost level of their application. Similarly, IT Architect (I10) mentioned how they are missing visibility to the costs. I10 also emphasized how this is a risk, as anyone could adjust the environment in a manner that accumulates major costs without knowing:

We have no visibility to the costs, we always must ask for them, and that is a challenge. We should have constant visibility, as anyone who has rights to the RGs environment can make changes. – IT Architect (I10)

Visibility has also been an issue for IT Architect (I4):

I would also add a general comment for all of it, maybe I shouldn't but still. So far because everything that has been built on the cloud foundation, we have never received an invoice yet, so that has never been a problem. – IT Architect (I4)

On the other hand, for Manager, IT & Digitalization (I11) and a few other interviewees, visibility is not an issue. I11 mentioned how they are able to see how much the spend in Azure is. The differences in visibility root from the fact that some cloud journeys took place prior to the establishment of the cloud foundation in the case company. I11 for instance is the owner of their business areas Enterprise Azure account which was established before the cloud foundation. I11 has ownership of the Resource Groups (RG) and deployments therefore, I11 has full visibility to the resources. However, majority of the interviewees have begun their cloud journeys after the formation of the cloud foundation, and they seem to be the ones with the visibility related issues. The cloud service partner has created a tool for cost visibility, but so far it seems that Senior IT Architect

(I7) is the only one that has rights to this tool or has used it in general. Moreover, some of the interviewees that are currently not a part of the cloud foundation are considering migrating their resources there in the future.

5.3.4 Problems with Capacity Management in the Cloud

As previously mentioned, many of the cloud journeys are still either in the very beginning or currently in the migration phase. For this reason, very few of the interviewees could comment on the problems related to capacity management in the cloud. Some interviewees that were already in the cloud did not have any specific issues however, a few interviewees could identify problems that have come up so far:

> *For instance we still had some problem which is under investigation, why there are no tags in some of our builds because that's how you understand what you are charged for, when you have tags, and is this this application, that application and so on, and right now we have a problem that in big part of our build there is no tag… But this is essentially the most important, it looks like you know tagging a picture or whatever, it looks funny but when it comes to the bill if you get 15k euro with no tag then it's like oh…something is wrong. – Senior IT Architect (I7)*

> *Sometimes we have issues with the storage account it might happen that it gets too many requests and we had some issues with the storage accounts not being able to handle that. – IT Architect (I5)*

> *Maybe a reoccurring problem is that when the database grows it gradually requires more capacity. In cases like these every now and then we need to check and maybe add some more capacity. – Manager, IT & Digitalization (I11)*

Senior IT Architect (I7) further mentioned the complexities of cloud bills:

> *This is really complex, it's not straightforward...for instance they are charging you for all the requests, post, get etc. and the charge is really small, because it's like 0,000034 cents, but then you look at the bill, and why this bill is showing 800 euros per month if this is so tiny. And then you look at the amount and it's like 60 million requests... And then there is some data transfer between the availability zones, and this is also going into giga bytes so you're wondering ok what is this. So generally, all the sections are kind of a bit cryptic and not really easy to read. When I started to discuss it with the vendor, they have dedicated teams who are working with… billing, and like you name it capacity management it can be kind of understood from a technical perspective is this instance doing its job is it running fine, but it can also be understood like ok is this instance too expensive… so I am*

referring more to the bill part. And this is then complicated. Although we are in a simple situation where we are not using serverless instances, the serverless they are notorious for being difficult to understand. – Senior IT Architect (I7)

5.3.5 Exit Plan

When asked about an exit plan for the cloud deployments, majority of the interviewees did not have any specific plan on what they would do if costs began to rise. However, most of the interviewees had thought about it to some extent. IT Service Owner (I6) mentioned that it should be taken into consideration during the planning phase. Some interviewees mentioned that an exit plan should be in place regardless of the costs, as certain other factors may lead to a need to exit:

You should think of an exit plan regardless of whether the costs are rising or not. You should be prepared for it due to all sorts of topics that could come, regulations that could come, simply new needs and other things, so sometimes you simply legally need to do something. Yes, it might cost you more, but you simply have to. – IT Architect (I4)

One thing that became apparent throughout the interviews was that the service models play a role in how easy it is to exit the cloud. Interviewees with IaaS deployments seemed assured that the servers could just be moved back to an on-premise environment:

So, for the IaaS of course any server is ok... if on a bigger case company scale, we somehow think clouds are costly and we want to go back to the on-premise data center, we can of course migrate our software. I mean we started from the on-premise data center, we migrated to IaaS, we can of course go back. – IT Architect (I4)

Move everything back to the on-premise data center. – IT Architect (I1)

IT Service Owner (I8) also agreed that most likely if costs were to rise uncontrollably, then maybe things will go back to the on-premise environment. I8 however, made a great point related to the scalability advantages of the cloud and how they could be enforced more extensively:

If the costs would rise uncontrollably… would everything go back, if it's cheaper to have on-premise… The advantages of scalability are evident for the larger units… perhaps stricter conversations and instructions to all the business users, that should the pay as you go model be used, is this needed on the weekends for instance, or at night and so on. Could we further limit the time that they are in use. – IT Service Owner (I8)

For IaaS service models however, reserved instances can become an obstacle when considering an exit plan. During Senior IT Architect (I7)'s cloud journey, a 3-year Reserved Instance (RI) commitment was made, as the costs were much lower than a 1-year commitment. For this reason, there is currently no exit plan. IT Service Owner (I9) further emphasized that if costs increased beyond their TCO, then the capacity should be moved to some other use within the case company:

> *So, the case company is a pretty large organization and as you know in cloud, we are just purchasing the capacity. It is not linked to any system so that capacity can be reused within the case company in any business group. So, I think it should not be a challenge in that way that we have been incurring the cost without someone using it. – IT Service Owner (I9)*

In addition, IT Architect (I4) also mentioned the possibility of shifting to a similar SaaS solution that is offered. However, the SaaS solution in this particular example was from the same provider, so in certain situations it would not necessarily guarantee a decrease in costs.

In comparison to IaaS deployments, interviewees with PaaS deployments had a mutual consensus that moving PaaS applications to another environment is not a simple task. For instance, IT Architect (I2) emphasized how it would be an extensive, expensive and difficult activity to recode cloud native solutions etc. as they are very heavily reliant on Microsoft products. Manager, IT & Digitalization (I11) agreed that for PaaS solutions exit plans are not easy.

IT Architect (I10) on the other hand, made sure that vendor lock-in would not become an issue. Instead of using i.e. Azure specific components, I10 used an SQL database that for instance AWS also provides:

> *For us, a cloud transfer for the application from place A to place B in the future has been done so that it was already thought about in the build phase. – IT Architect (I10)*

However, I10 further stated how an issue that arises from a homemade solution is the ability to keep up with constantly changing technologies and other things. According to I10, this is a risk that needs to then be taken.

Furthermore, Senior Manager, IT (I12) has a large-scale exit plan, which involves shifting from Azure to AWS. In addition, the current platform has been designed in a way that it includes many different Azure services, so if one service begins to cost too much, they can flexibly change it to a corresponding component. IT Architect (I5) similarly contemplated on different options such as moving back on-premise or switching cloud providers.

In addition, I5 pointed out another relevant fact related to moving everything back to an on-premise environment:

> *We don't have anything on paper, but if costs begin to rise, we could bring everything on-premise, but that will also have a cost impact because we are going to need to buy hardware so. It's quite difficult this exit plan when of course you can try to see other vendors like Amazon or Google, but we are really comfortable with Azure and how they provide this it's quite good for Microsoft developers. IT Architect (I5)*

5.3.6 Roles and Responsibilities

The roles and responsibilities for planning capacity and the capacity design varied. Certain interviewees conducted this on their own, whereas others used vendors to help and provide these details. IT Service Owner (I6) summed up how planning for capacity and the capacity design could be conducted:

> *If user amounts affect the capacity, then I should find those out. Service owners need to clarify this type of need from the business and then inform this to the individuals who specifically define the capacity. In my opinion the capacity needs should come from the application vendor, as in what types of servers, how much disk space etc. Also defining the usage level of the system. It should be the service owners' task to clarify this from the business. – IT Service Owner (I6)*

The main pattern among all interviewees was that this activity needs to be done by someone who has the appropriate technical knowledge of the application in question:

> *We cannot assume that individuals from the business without technical knowledge can estimate the capacity need. – IT Architect (I2)*

The interviewees were asked about who should be responsible for following the capacity related details. Similarly, to the capacity planning and design, some interviewees either conducted or believed that this activity was their responsibility. On the other hand, others used vendors to help with the activity, or believed that this should be fully handled by the vendor or cloud service partner.

The interviewees were also asked for their opinions on the relevant individuals needed to make any type of change that comes up related to the capacity. Majority identified how the vendor and cloud service partner should apply the changes. The approval and identified needs however, should come from the case company. In addition, the vendor could also suggest needed changes. IT Architect (I4) emphasized how the change should be performed by the vendor, but the design and general things need to follow the guidelines

of the case company. In other words, the case company should approve the change. Other interviewees agreed with this perspective:

> *And I guess the process goes like this that the case company is requesting the officially change, then the vendor or in a way vendor is suggesting the change, and then I'm approving it and then the final purchase is done by the cloud service partner. – Senior IT Architect (I7)*

> *We identify and we tell the need that ok, this is when the system needs to be switched off, or this is when we need the computing back and the vendor then performs those actions together with the cloud service partner. – IT Service Owner (I9)*

> *It would be good if the vendor gave recommendations, such as now it looks like you have too much capacity, have you thought about decreasing it… The person that then decides on it in the case company needs to be someone who understands it… I would say the service manager together with the IT architect. – Senior Manager, IT (I14)*

Senior Manager, IT (I12) also mentioned how parameters could be set to avoid having the vendor ask permission for every single change:

> *We would most likely give parameters to the vendor, so then when the limit is exceeded, make the change in a way that they do not always ask us approval for every single thing. – Senior Manager, IT (I12)*

5.4 Cost Optimization

The interviewees were also asked about cost optimization related topics. The interviews included discussion on the importance and motivation towards optimizing costs, cost optimization methods currently used, lessons learned, when cost optimization is considered worth it, licenses and expectations. Table 4 summarizes the following sub-sections by listing the topic in question, a short description of the content and either a direct citation or short description of the results.

Table 4. *Summary of the cost optimization interview results*

Topic	Description	Results
Motivation	Motivation to optimize cloud costs	*I argue that cost optimization has always been a central topic in the case company - (I2)*
Methods	Cost optimization methods used for cloud applications	Service models, cost models, tradeoffs between the different options and other cost optimization related factors
Lessons Learned	Lessons learned so far in the cloud journey	*Purely our design, you can go in a thousand different ways, but with this flexibility comes the burden of responsibility that you then need to decide which design is technically optimal, but at the same time what would be the cost of it - (I7)*
Worth	Opinions on the worth of cost optimization	The work itself should cost less than the gained savings
Licenses	License related costs, thoughts and decisions	*There's no general way of doing it. You just need to dig in the details and do the calculations and kind of see different options and how much it would be - (I4)*
Expectations	Expectations related to cost optimization	Improved visibility, consolidation of resources, consultation, tools…

5.4.1 Motivation to Optimize Costs

Prior to moving to the cloud, employees of the case company at all stages of the cloud journey have identified the need to consider costs. Clarifying costs to the business is a key starting point and having the ability to estimate run costs is highlighted by IT Service Owner (I6). Furthermore, IT Architect (I4) emphasizes the importance of keeping costs

low, as the different business areas expect and anticipate costs to be low. The case company and its different business areas are all cost aware:

> *We are very cost aware, and our business is cost aware and demands that the cost details are clarified before we go anywhere. – IT Service Owner (I6)*

Similarly, to cost awareness, cost optimization has been a topic long before the cloud journey:

> *We are always trying to optimize the costs. – IT Architect (I5)*

> *I argue that cost optimization has always been a central topic in the case company. It has been in the past, but more related to buying one or two or three servers. We have had to invest in the hardware… However, scaling up and down has not been very agile. – IT Architect (I2)*

Although the need to identify costs is consistent throughout the case company, there is no clear unified process, as certain interviewees have faced major challenges with getting started. This is especially apparent for individuals who are in the planning phase of the cloud journey:

> *It is not clear, I do not know if it is clear for the case company. Not getting the costs has been an ongoing issue…. Maybe this is still new to the case company, there are no readymade processes that state do this and do that, we are still in the learning phase. We are moving forward according to feel, and whenever questions come up, we begin to search where we could find the appropriate answer. – IT Architect (I1)*

> *This cloud topic is very new to me, so understanding what all needs to be taken into consideration in the beginning is a challenge. – IT Service Owner (I6)*

All the interviewees also agreed that cost optimization is very important when the application is in the cloud. The overall motivation towards cost optimization was evident. Project Manager, IT (I13) specifically mentioned that they are at a phase in their project where cost optimization is pivotal:

> *We have a very limited budget for running this and everyone of course wishes that savings can be achieved… We had a certain amount of infra budgeted for the business case, but of course if we can optimize some 10 000s off a year, that would be great. – Project Manager, IT (I13)*

5.4.2 Cost Optimization Methods

The importance of cost optimization was evident to all the interviewees. Senior Manager, IT (I12) emphasized how optimization is important, as I12 wants to ensure that they are only paying for what is being used. Depending on the phase of the cloud journey, the types of methods used to optimize costs varied. A few interviewees stated the importance of having a good idea on cost optimization at the very beginning of the cloud journey. For instance, IT Service Owner (I9) mentioned how already during the planning phase certain things were known that would change from on-premise to the cloud. They had six test systems that were not needed throughout the year, but the on-premise set up did not allow reductions in capacity or switching things off when not needed. In addition, IT Architect (I10) took cost optimization into consideration from the very get go of the project. I10 further designed the environment in a way that takes future projects and applications into consideration from a cost optimization perspective:

> *The cloud pushes to optimize costs, as the costs are all available. – IT Architect (I10)*

Senior Manager, IT (I14) believed that the extent of cost optimization is partially dependent on the targeted solution. When talking about SaaS services, the available options are basically all in the cloud. The choice itself will depend on a combination of the suitability to the business process versus its costs. I14 further highlighted, when considering options other than SaaS, there may be more factors to choose from. Options range from on-premise to cloud solutions, which include i.e. infra, platforms and so on. I14 mentioned that in these types of cases there is most likely more focus on cost optimization itself.

The different service models offer varying opportunities when entering the cloud environment. This was something that Senior IT Architect (I7) realized during the cloud journey. According to I7 they started with the rehost model however, after calculating costs and making estimations they realized that they are better off revising:

> *So, then that was one of the problems like how to design it in the most optimal way. We are going from the on-premise to the cloud we don't want to make too many changes to our landscape, but at the same time we want to have the cloud flexibility and optimize the costs and meet all these targets. – Senior IT Architect (I7)*

The original plan was to just lift and shift the instances to AWS. The turning point was when I7 realized that some of the instances were rarely used, and therefore they did not need to be RIs and could instead run on-demand:

> *But then the trick is that on-demand is a couple of times more expensive so then you need to shut them down and then agree with all stakeholders that they will be down and the moment you want them up then you need to request it. – Senior IT Architect (I7)*

Furthermore, I7 emphasized how capacity is directly related to the design. They built three different scenarios for the design in order to assess the impact on capacity. This was done to check the costs, as capacity more or less remained the same, but the cost was very different:

> *On-premise for instance we have 1 server which has 3 test environments hosted to limit the number of servers and so on, and then when we did this exercise, we realized that in fact it would be more beneficial to put every test box in a separate instance and the reason for that is that then you get more flexibility… If you have 3 servers in a huge instance, then you need to pay for it all the time, but if you have 3 servers in 3 separate instances you can shutdown 2 and run with 1 all the time. – Senior IT Architect (I7)*

Project Manager, IT (I13) as well as other interviewees made several calculations of the possibilities in the cloud. For instance, the differences in price for RIs, on-demand, on/ off were checked. The costs became very evident to I7, and for that reason certain changes were made. I7 had things running on-demand to begin with. Shortly after they were changed to RIs, as there was a huge difference in price:

> *If you have some peanuts, then it's no brainer you can first run a bit on-demand and go to the RI, but if you have something like this so this Sidecar Hana it's something like 20 dollars/hour, so this is generating thousands of euros within weeks. The only instance which we didn't take as RIs were the ones which we decided that they will not be in use or they will be in use but for couple of hours monthly, so then it doesn't make sense to take RI. – Senior IT Architect (I7)*

Furthermore, the type of application at hand affected the cost optimization methods used. Several interviewees discussed the need to understand the nature of the application in order to know what works for it best. IT Service Owner (I8) mentioned the pay as you go possibilities in the cloud, such as optimizing costs outside of business hours. I3 further mentioned that some servers have been tagged according to shut down schedules i.e. night time. However, for I8 the pay as you go model was not an option, as the application is needed 24/7. Starting small and scaling was something that I8 could do to achieve cost optimization. IT Service Owner (I6) on the other hand, is planning on moving a database archive to the cloud. I6 was still in the very beginning of the cloud journey but

identified that bringing the servers up and shutting them down must be configured, as the servers should only be running when they are used. Project Manager, IT (I13) mentioned how currently the cloud service partner shuts the servers off or changes the capacity upon requests. For now, according to I13 this is the only way of doing this however, I13 is hoping that in the future the on/ off functionalities could be done using a script, or via a button that has this functionality in it. Moreover, IT Architect (I4) discussed how when the instances are stable and running 24/7, then RIs are a good choice. On the other hand, for instances that are only required during business hours, auto shut downs should be put in place. Senior IT Architect (I7) summed up a process that is taken to check the appropriate cost model:

> *So, from our perspective we always approach it in this way that is it really needed, and then if its needed then do we take it on-demand, RI 1 year or RI 3 years. – Senior IT Architect (I7)*

IT Service Owner (I9) also mentioned how the different cost models were utilized according to the environment:

> *During the project phase we had taken most of the systems as 3-year RI capacity, which means for next 3 years we cannot actually change anything, and we don't want to change because our environment is actually pretty stable… There are only 2 test systems which are on-demand, which means we have the option to turn it off when not needed so in that way we are bringing capacity improvements to the run. – IT Service Owner (I9)*

A lot was discussed on the different options, and interviewees had fairly good ideas on what could be done cost optimization wise according to the nature of the application. Project Manager, IT (I13) however, mentioned how the different options may need to be weighed against each other:

> *I wonder about the on/ off capabilities, would that be a good solution for us…If we get to the same result by locking the resources at some certain level, then that might be a more beneficial solution for us. – Project Manager, IT (I13)*

In order to be able to decide on the correct option, it became apparent that there may need to be a certain time period that the application is monitored before the decision is made. For some interviewees the obvious choice has been evident moving from on-premise to the cloud however, in certain situations the decision may not have been made prior to moving to the cloud. I13 emphasized how there needs to be a monitoring period, and an overall outlook on the environment to know whether i.e. less resources could be

used. I13 further mentioned how a lot can be saved by just following the CPU and RAM usage and checking if using less resources is feasible:

> *Could we survive with less resources, that requires trying it out and calculating the capacity and then seeing if the regular usage is still working, or have issues come up. – Project Manager, IT (I13)*

Senior Manager, IT (I12) further discussed how their area has a large amount of different services, and capacity, so each one is managed in a slightly different manner:

> *Majority of them are autoscaling… we have parameterized them in a certain way, as in how much they can scale and how fast etc. Then we have some applications and services where we have to make the decisions and changes as we go. For instance, the small amount of IaaS that we have, we have to adjust them manually. Then for example we have some serverless things, such as Azure functions applications where we don't have to care about it at all. – Senior Manager, IT (I12)*

For IT Architect (I2) the resources are fairly automatic and pay as you go is used. They have functions and API management which cost according to the use i.e. the amount of calls and the number of seconds that they are running for. Furthermore, IT Architect (I5) uses autoscaling to manage the capacity. I5 further mentioned how autoscaling may at times point out the need to fine tune:

> *If you are using the standard which is the cheapest one there is not much room to optimize the cost there, but if this is autoscaling means that we are having more costs, so what is happening. And usually what comes up from that is that we need to do some corrections on the code and the application stabilizes. – IT Architect (I5)*

Chief Architect (I3) further mentioned how their development environment (dev, test, prod) needs to be rechecked. The dev and test could run at a lower performance level. Therefore, I3 wants to ensure that they are running at a level that is cost optimized, especially since new services have been added. In addition, I3 mentioned how for services that are new, in the beginning it would be good to gather some experience of the way the service functions. Once enough experience has been gathered, the services could be automized for i.e. services that allow hourly adjustments. But once again the cost of the automized service would need to be weighed against the potential savings. The development of the automized service would need to be coded with an Azure function so that it checks the metrics and usage and then scales. This cannot be considered a simple task.

In addition, Manager, IT & Digitalization (I11) mentioned the importance of checking for newer versions and functionalities in the cloud, as this could entail better functionalities at lower costs:

> *But of course, the technologies change and so on, so sometimes there is the chance to lower capacity when things are done in a smarter way and even change database types, as Microsoft has brought more features from the premium to the standard and so on. – Manager, IT & Digitalization (I11)*

Furthermore, IT Architect (I10) talked about the fact that a lot of the time Azure pricing might not depend on needing more capacity, but instead it will depend on some functionality that requires an increase in capacity.

5.4.3 Lessons Learned

There were a variety of different lessons learned throughout the cloud journeys. The extent of the lessons depended on the phase of the cloud journey.

IT Service Owner (I6) was still in the very beginning, planning the cloud journey. According to I6 it is not really easy to get started, as there is no specific bible in one place which would give guidelines on how to start and get a good end result. I6 further mentioned the complexity of the cloud:

> *These cloud related things, it feels like they can be done in so many ways, so these first cases in my opinion are very educational for understanding the bigger picture. – IT Service Owner (I6)*

Senior IT Architect (I7) similarly mentioned the large amount of options in the cloud and emphasized the importance of really taking time at the beginning to understand the design and costs:

> *Purely our design, you can go in a thousand different ways, but with this flexibility comes the burden of responsibility that you then need to decide which design is technically optimal, but at the same time what would be the cost of it. – Senior IT Architect (I7)*

IT Architect (I2) emphasized the importance of using smart modern technologies. I2 gave an example on how this year they are going to move a SQL database to the cloud to get rid of the on-premise server. DBaaS will be used and refactoring will take place for the migrated components, as they do not want any database server. I2 places effort on designing:

> *An important thing is to not only do a "lift and shift" from the on-premise data center to the cloud, but instead think about things, refactor and look at the architecture. – IT Architect (I2)*

When IT Architect (I5) started the cloud journey around four years ago, there were not many people with knowledge on Azure. The cloud was very talked about however, it was a little difficult to execute and took many trials and errors. I5 sees this aspect of the cloud positive, as things can be tried and then shut down if needed and then done differently:

> *But the knowledge was a key topic and I still think it is because if you try to keep up with the technology its always difficult to find people that can actually keep up the pace and in the cloud it's all about new technology so it's difficult to follow up. – IT Architect (I5)*

I14 mentioned that the sales argument about price elasticity is good however, it requires a lot of work to achieve. I14 wondered how well do the prices in practice scale downwards. From a SaaS point of view, paying a certain amount per month per user with a price that is elastic requires very active user management. I14 mentioned how the price itself will not decrease if all the users are still being charged the per month price, even if they are not using the software.

According to Senior Manager, IT (I12), the case company's internal IT processes and process management models are not suitable for platform thinking and the switching of components. Agile methods, such as taking infrastructure and components into use, removing them, moving them around, are very hard to adapt into a rigid organization using Information Technology Infrastructure Library (ITIL). I12 mentioned that this is a cultural challenge that is currently very evident within the case company.

IT Architect (I4) emphasized the contractual part, access management more specifically. Having two components on the same account, that possibly have separate Application Management Services (AMS) partners requires ensuring that the AMS partner has access to the correct component. In addition, from the case company's point of view, what should go via the case company for approval and what should not. In the cloud when someone is given an account basically they are given a full on-premise data center of their own. The user could manage network settings, application settings, and they could spin off servers:

> *So how do you model it… now suddenly you have one team that could do anything, and they could do it very quickly, so control points basically. Access management one of them, knowing the environment and what is there, the fact that it follows our guidelines and the direction, still learning. – IT Architect (I4)*

Senior IT Architect (I7) further mentioned that they did not have issues with the technical factors, but instead struggled more with the costs:

> *It is quite premature to say that it will be in a range of 100k or 50k or 150k. You really need to do your homework, calculate everything and then you will see, ok this is it we might make some rough assumptions we might make some small or bigger mistakes, but now this is the basis for us to say that this is our capacity need and this is the price tag and at the end of the implementation it might differ by 10%, 20%, but you have some kind of baseline. You just cannot say that, yeh it will be more or less 100k. – Senior IT Architect (I7)*

Manager, IT & Digitalization (I11) discussed how it is always important to really question case by case when on-premise solutions are being taken to the cloud, to understand the benefits that could be reaped from changing the environment. I11 gave a specific example where on-premise still makes more sense than the cloud. They have solutions which require very heavy optimization, which would cost a lot more in the cloud, so for this reason these types of solutions have still been kept on-premise. The use and the load may be factors that affect this. I11 mentioned however, to reach a similar price in the cloud and on-premise, extensive optimization work would need to be done, such as running instances only when they are in use and turning them off when they are not used.

5.4.4 When Cost Optimization is Seen as Worth it

The main opinion on when cost optimization is seen as worth it was that the work put into the optimization should cost less than the gained savings of the change. If it costs a lot more than the benefits achieved, then there is no point. IT Architect (I10) mentioned that getting money back, simplification, better management and supporting the vision are all factors that could be worth it. According to IT Service Owner (I9) it may even take a couple of years to see if the optimization itself was worth it. I10 similarly stated:

> *Of course it is when we get the cost back in the next three years from making the change. – IT Architect (I10)*

Senior IT Architect (I7) agreed that in order to see savings, work must be put into it:

> *I think you know to put it into a figure maybe if it's over 10k per year, but then in order to find out what would be the cost saving then you need to do this exercise. – Senior IT Architect (I7)*

A few interviewees contemplated on whether small amounts should maybe be overlooked however, majority concluded that they do in fact accumulate large savings, depending on how the situation is looked at:

> *For a few tens it's not worth it, then rather just accept the costs. But then looking at the bigger picture if you have many hundreds it will become thousands. But who watches over this, I look at a narrow sector, someone should look at the bigger picture. – IT Architect (I1)*

> *Even a small saving is a saving… If we have a lot of servers, then the savings would most likely be very large in the long run… I believe that even a small saving is a saving, as they all accumulate. – Project Manager, IT (I13)*

> *A basic investment way of thinking, that if with a little effort you get a reasonable benefit, then it is worth it, because often we go after the big ones so that huge savings can be gained, but oftentimes the big savings build up from the smaller ones. – IT Service Owner (I8)*

IT Service Owner (I6) emphasized the time factor in the optimization activity:

> *The benefits achieved from the work need to be large enough to at least cover the amount of time it takes to understand and perform the change. – IT Service Owner (I6)*

Senior Manager, IT (I12) focuses on things that require large amounts of capacity. Bigger masses are prioritized over i.e. individual calls. However, I12 continued that autoscaling takes care of the smaller things efficiently. Similarly, IT Architect (I5) emphasized the benefits of autoscaling:

> *I think it's always worth it and what I see with this autoscaling is that you can work with lower costs, and if needed you can increase those. So, I think that actually helps us quite a lot, so we don't need to be over assigning resources to anything. – IT Architect (I5)*

The design itself was once again mentioned. IT Architect (I4) stated:

> *So ok, the correct answer, or the one we would aim for is that the design of the application takes the costs into account. So, then you shouldn't need to kind of every time think about this separately, but its rather incorporated into the design of the application. – IT Architect (I4)*

A very relevant point was added by Senior Manager, IT (I14). A risk factor should be taken into consideration:

> *For instance, the capacity management or cost optimization activity should not affect the business. It should somehow be made sure that the decisions made do not cause problems or risk the business. – Senior Manager, IT (I14)*

5.4.5 Licenses

Licenses varied a lot among the interviewees. IT Service Owner (I9) mentioned how licenses were discussed throughout the project:

> *We have been thinking what type of licenses are most suitable to run this type of application on AWS. That means do we take the native AWS licenses for this or should we purchase xyz from the AWS marketplace. So, I would say all the time this has been discussed throughout the project and also, I think it will be discussed in future as well, when these licenses are going to expire. – IT Service Owner (I9)*

Senior Manager, IT (I14) mentioned how SaaS services most likely have less license models. The optimization roots from user roles and the extent of the user rights i.e. super user's versus normal users. Moreover, IT Architect (I4) summed up how licenses truly depend on the chosen deployment model and tools:

> *It depends on the tool we are talking, so if we talk about PaaS lambda functions there's no license cost... When we talk about IaaS that's definitely no different than physical servers. The calculation may be different in the cloud. – IT Architect (I4)*

> *When we have EC2 and we install our own application, so that of course has the license cost and that has been included in the calculation… But that goes so deeply to the application itself that different vendors have different ways to calculate the license, but there's no general way of doing it. You just need to dig in the details and do the calculations and kind of see different options and how much it would be. – IT Architect (I4)*

IT Architect (I4) further mentioned how the different service models in the cloud may need to be compared against each other when it comes to licenses and agility:

> *One of the platforms we have as a PaaS, it is now heavily discussed that should it be SaaS, so the provider is offering that also as SaaS. It is license costs and general costs versus agility... When you own it you of course could finetune it and configure it. When its SaaS you cannot so, they are always … in kind of need to find the right balance between all of them. So, you might need to pay more but if you gain agility that is important, maybe that's the right choice. Definitely the main thing is that it doesn't come as a surprise, so it needs to be a decided choice not a retrospectively kind of oh it cost more. – IT Architect (I4)*

IT Architect (I10) emphasized how licenses and the use of licenses were extremely important. User based licenses were mentioned by several interviewees. Some solutions

require user-based licenses therefore, a fee is incurred whether the user uses the service or not. For this reason, I10 chose a solution that was based on the usage of resources.

Senior IT Architect (I7) on the other hand took the pay as you go license model, and did not carry out any specific license optimization exercise:

> *We essentially decided that we take pay as you go, those licenses for whatever we can. There are only two things which we had to buy separately so far, first was this Hana DB license from SAP, which is like bring your own license, and then the second was this EMC networker for backup… We didn't have any exercises for optimizing them. – Senior IT Architect (I7)*

Chief Architect (I3) further emphasized how so far, they have not had to think about licenses in the cloud. For PaaS services the cost of the license is included in the price:

> *When you don't have servers, you don't have to think about OS licenses and other things, so that's been pretty handy. – Chief Architect (I3)*

In Chief Architect (I3)'s case, the system is developed using .NET. However, the developers need Visual Studio and other developer licenses. I3 also talked about the importance of analyzing the entire stack especially when custom applications are being made, in order to prevent license related problems. Senior Manager, IT (I12) for instance incurs very few license related costs or limits on the cloud platform. However, SAP or other outside data sources may restrict the use of data on a cloud platform. In other words, the number of individuals that can use the data may be limited. Furthermore, IT Architect (I2) mentioned how they use i.e. Azure DevOps, and always try to use the cheapest licenses as well as limit license rights for the users. I2 further informed that they try to use as few licenses as possible and check them to ensure that they are being used. I2 stated that typically checks are done on a yearly basis once a license has been assigned to a user.

5.4.6 Expectations

The interviewees were asked for opinions on the what they expect regarding cost optimization from the case company:

> *I hope finding and clarifying costs will be simpler in the future. Either information on how to find the details or some kind of help, so that it would be easier to get started. So far it feels like we have taken a lot of time and I have tried to ask and understand the costs and that has taken a lot of time. – IT Service Owner (I6)*

IT Architect (I10) emphasized the importance of cost visibility and suggested how a Power BI report could be used. Similarly, IT Architect (I4) mentioned visibility as an important factor:

> *Definitely visibility, so you know some dashboard or something that you could ad hoc login… But at least weekly or something that what is the spending, what was it spent on, so that we could do our own analysis and kind of know which component is the one that is costing. Because if only at the end of the year you suddenly then get an invoice it's very difficult to then know what happened in the meanwhile. – IT Architect (I4)*

Furthermore, a tool that could be developed in a simple manner to optimize larger cloud workloads was suggested by Chief Architect (I3). For instance, autoscaling a larger volume of resources according to the workload.

Senior Manager, IT (I12) mentioned how it would be nice if someone could take care of the IaaS optimization and have some kind of centralized cost and capacity management. This would enable I12 to focus primarily on the development of applications. IT Architects (I1) and (I4) and Senior IT Architect (I7) also mentioned centralized services:

> *Maybe not for one application, but for the whole organization, someone that would look at the overall situation. If something were to be done what would its affect be. – IT Architect (I1)*

> *On a very high level and I mean again I don't know what is feasible and I also know the case company's internal structure is such, that it is very difficult to achieve such things. But the best from my point of view, would be that someone does a full optimization… You see that my EC2 are statistically running for those hours, so you buy the reserved capacity for it and then it costs me less… I also understand that it cannot be that someone who does not know the application suddenly makes decisions on it... So based on statistical historical usage offering, what could be done and maybe in that regards also informing the options… We know about instance reservations, but there might be things you could do that simply it doesn't make sense for every team to read about all the options themselves and think about them. So, if someone could be an expert of optimizations and suggest what you could maybe do then it's easier for the application teams to say, yeh that makes sense or no in my case I know that tomorrow the capacity will be double, so no need to do it now etc. – IT Architect (I4)*

> *I see the reason why we could utilize a centralized service for instance, if I'm buying let's say 20 instances which have this and that capacity and then you're some*

> *other application, there is some other application and all together we are buying already like 50 instances, so maybe there is some reasoning here that they should be purchased as a bundle and that will give some additional discount. But then the tricky part with this centralized service is that would they really understand the individual application need, or will they be more from the commercial perspective. The application would need to say that they need this and that and then the centralized service will say ok, so this is your need we collect all the needs from everyone and then we make some better deal for you. Ideally, they could help us as well with these individual instances, but here comes the specialization problem, so will they be specialized in all these applications, will they understand them I don't know, this is a bit tricky. – Senior IT Architect (I7)*

IT Architect (I5) similarly discussed the possibilities of using resources more effectively and in an optimized manner, as well as combining resources:

> *Currently we have different resource groups for each application, so each application has its own resources and its divided there. But now with this new initiative, the idea is to not have one resource group per application but start consolidating and make better use of the resources. So maybe it's more cost effective to have one SQL server with different databases and then you share those with different applications, than have just one small SQL server per each. So, SQL server is just an example, but I think that this, how to effectively use and share the resources among different applications is something that is good to have. It makes it harder to follow, but you might have more benefits there on the costs. Of course if you have a resource group dedicated to an application then you know that if something happens then it's this application, but if you start sharing then ok you need to try to find who is impacting the most and you need to be sure that this application that is impacting there is not also making the others feel the downtime... So, you can have cost benefits there and we are trying to see how we can achieve that, but we also need to be careful, so that one application is not penalizing the others because of bad coding etc. But it already happened that the vendor did not develop some functionality well, and it was consuming a lot so how we can ensure that this is not affecting others if you are sharing those resources. So, it's just a matter of analyzing that but I think that it might be good. – IT Architect (I5)*

Senior Manager, IT (I14) split the different service models. For SaaS I14 mentioned that a centralized service could be difficult. For individual SaaS cases, need basis help could

be a good fit. For case companywide platforms, a centralized service could provide efficient service and assistance. For instance, if several different business areas ERPs were on AWS, the volume would be big enough to have a centralized service for the platform's capacity management. In other words, the centralized service would serve the business IT services within the case company. I14 summarized:

> *For SaaS services assistance on a need basis and then from the platform perspective a centralized service when the volume is big enough. – Senior Manager, IT (I14)*

Consultation was also brought up by several interviewees. IT Architect (I2) for instance mentioned how it would be very nice to receive some advice on their resources and whether they should be replaced with something once they migrate. Then a revisit after 1 or 3 or 6 months when data is available to check the usage and spend and see if something could be optimized. I2 further mentioned that down the line it could turn into a once a year type of activity. Furthermore, Project Manager, IT (I13) discussed how a regular audit like check-up would be good to ensure best practices are used and costs are optimized. In addition, I11 mentioned how a case by case check should be done from a cost optimization audit point of view. Senior IT Architect (I7) similarly mentioned:

> *Yeh I think we should be advised in a wizard way how do you build your kind of capacity requirement, so step one you do this, step two you do this and so on… We were moving in the fog at the beginning. One thing which was really underestimated is that this is a simple activity, just put few lines in excel it will calculate for you and that's it, but it's not. First of all of course you start from your existing environment so you need to think… so this should be the first thing do you want to revise, rehost or do it as a PaaS or SaaS or this, so this is the first thing and then maybe you do some kind of circle that you go back after, but yeh would be really great to have some kind of guidelines how to do this exercise. – Senior IT Architect (I7)*

> *If you want to really move your application to the cloud, then first you need to look at the big picture and then down the line during the process you fine tune it, so you are saying that ok, now I see that this is not performing well, or now I see that this is maybe too big of a machine, so we should downscale. So, then this should come as a kind of optimization proposal for finetuning. But probably the more time you spend right at the beginning to assess your needs and then build your design, then less of a finetuning you need later on. However, this is probably unavoidable,*

> *that you will discover something, some dependencies or some new things which will change your demand down the line. – Senior IT Architect (I7)*

The speed of the technological advancements was something that IT Service Owner (I8) mentioned. I8 further discussed how it is very difficult to keep up with these advancements and to know what the available options are for different types of deployments. Therefore, someone should know how to ask the correct questions at the beginning of the cloud journey to ensure that the clouds full potential is reaped from a cost optimization perspective:

> *The questions would need to be precise, such as is it used on the weekends. – IT Service Owner (I8)*

IT Architect (I5) continued how keeping up to date with new ways of working would be good however, I5 further mentioned how the new technologies can be unpredictable:

> *Also seeing if there are new ways of working. Now there is this serverless which they say is cheaper. Last week I was doing some personal tests and something that should be two dollars ended up spending all my Azure credits that I have in two days. So, if I have this database it would cost in a month 15 dollars and supposedly this serverless would cost just 2 dollars. So, I say ok that's nice let's do that, and in two days I was able to spend all my credits. I think now we are in March, so I think I have my credits again. But you need to be careful with those because there are a lot of beautiful names there like serverless and do that do this. – IT Architect (I5)*

IT Service Owner (I9) further discussed how testing data should be minimized:

> *I definitely would like to see in the future for example when we are doing this testing you know in our supply chain environment, so there are three major business releases in a year and when we are doing the testing, we are actually doing the testing with the full scope production data. That is not actually something I would like to see in the future, so when we are having cloud, we should have some type of possibility to slice the data so that we are doing testing with only one year past data rather than 15- or 10-years data, which is not needed. We don't need that much capacity to do the testing for the releases. So, I think somehow as a service if someone provided test data as a service, or test as a service, or data as a service in that environment, I think that is going to be really great for future. – IT Service Owner (I9)*

6. DISCUSSION

In this chapter, findings from the empirical results and existing literature are discussed and the results are presented. Chapter 6.1 presents the overall process. Chapter 6.4 depicts a more specific process for the case company. The process is broken down into numbered activities and the numbered activities are further detailed in chapters 6.2 and 6.3. In addition, chapter 6.5 presents the recommendations for the case company.

6.1 Overall Cloud Journey Process

From the empirical results and literature review, several different cloud journey processes and sub-processes could be identified. The application cloud journey begins with the Business Case Process. Once the business case has been completed, the Prior to Cloud process begins. The process then moves to the In Cloud Process and ends with the Exit Cloud Process. Two sub-processes were further identified, the Cost Optimization & Capacity Management Process, Prior to Cloud and the Cost Optimization & Capacity Management Process, In Cloud. Figure 22 depicts the higher-level image of the cloud journey.

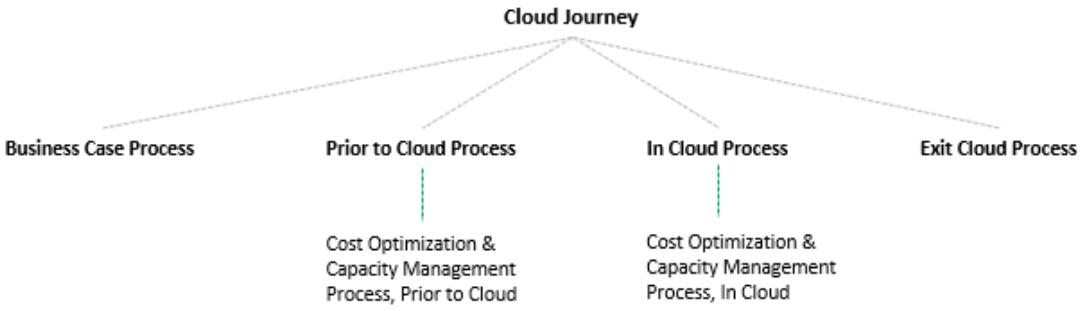


Figure 22. *Cloud journey processes & sub-processes*

Each process and sub-process consists of several activities. The relevant activities can be identified by combining findings from the theoretical background and the empirical results. The buildup of the processes, sub-processes and activities relevant to cost optimization and capacity management are further detailed in chapters 6.2 and 6.3.

The empirical results highlight how the cloud is constantly evolving and technological advancements are difficult to keep up with. Additionally, the results highlight how there are many options on the market, further complicating the decision making from a technical and cost optimal perspective. Literature similarly identifies how public cloud environments are extensively complex by nature (Mithani et al. 2010). In addition, finding the

most optimal solution according to application requirements while keeping cost-effectiveness in mind has proven to be difficult (Evangelinou et al. 2018; Huang et al. 2014). Literature also emphasizes how cost optimization should be an ongoing practice (Sabharwal & Wali 2013; Anderson 2018; Cristea 2017), which identifies and prioritizes cost optimization initiatives (Cristea 2017). In other words, the public cloud is constantly under change making it harder to find solutions for applications that are cost optimized to begin with, as well as remain cost optimized throughout the cloud journey. Therefore, in addition to the identified process in figure 22, a KM process must be established.

Literature emphasizes the importance of a cost-conscious organization (Microsoft Azure 2019). In order to achieve a cost aware cloud adoption, relevant stakeholders within an organization must be included (Amazon 2018), and an appropriate team for the cost optimization activities should be established (Cristea 2017). In addition, the importance of interaction and collaboration between internal and external stakeholders was evident from literature (Prasad et al. 2014; Willcocks et al 2013). Furthermore, literature states that specific qualities, knowledge and skillsets must be present to ensure that the cloud is used in an appropriate and cost optimal manner (Prasad & Green 2015; Willcocks et al. 2013; Marston et al. 2011). As an example, a thorough understanding of the cloud computing offerings is required in order to define the computing requirements (Willcocks et al. 2013). Literature further highlights how relevant qualities within organizations should be used instead of establishing completely new IT governance structures, to avoid unnecessary costs (Debreceny 2013). Empirical results similarly suggest that an understanding of the cloud offerings is essential. In addition, empirical results highlight how application specific knowledge is important. Therefore, internal employees of an organization and external vendors are required to perform the activities along the business process. Roles and responsibilities must also be assigned to the KM process (instruments, knowledge and tools).

Figure 23 depicts the combined business and KM process, as well as identifies the relevant parties:

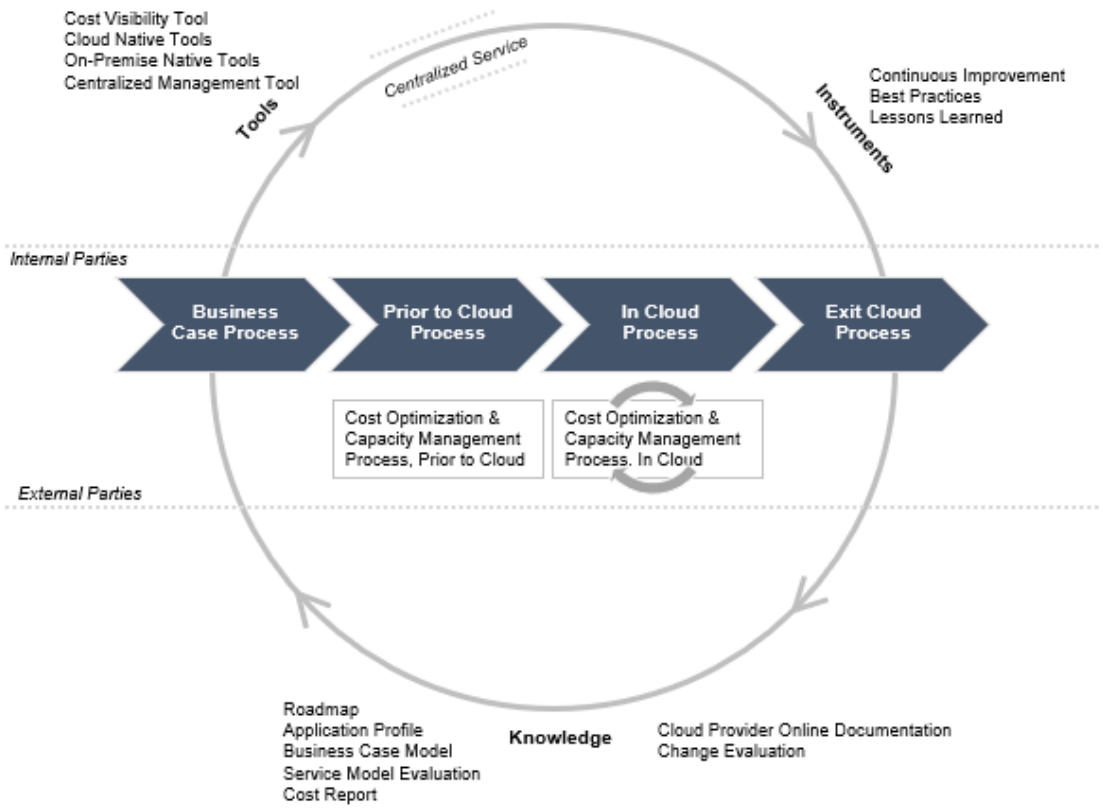


Figure 23. *Overall cloud journey process*

The KM process supports and ensures the business process is up to date with the rapidly changing public cloud environment. The instruments, knowledge and tools of the KM process, assist and guide with decision making, as well as performing activities along the business process. The relevant internal and external parties along the business process must understand the cloud and its offerings, as well as have application specific knowledge in order to optimize costs and manage capacity efficiently. The KM process must be centrally managed within the organization and enable cost optimization and capacity management. The KM process must ensure that the parties performing the activities along the business process have the needed support for a successful cloud journey. Figure 23 is depicted from the point of view of a cloud consumer organization.

6.2 Cloud Journey Phases

The cloud journey consists of four different steps as presented in the Cloud Journey Phases, Level 1 process drawing. The cloud journey begins with the Business Case Process. Preimeserger (2017) and Mithani et al. (2010) emphasize the importance of a business case prior to moving from on-premise to the cloud. Once the decision has been made to move to the cloud, the Prior to Cloud Process begins. The importance of spending time on the design prior to the cloud was evident from the empirical results. The In

Cloud Process begins once the application has been deployed in a public cloud environment. This process will last according to the contract period, or until a decision is made to exit the cloud environment.

6.3 Cloud Journey, Cost Optimization & Capacity Management

1. Initiation of Cloud Journey: The first activity of the Business Case phase entails the need to have a clear roadmap. The empirical results suggest that the roadmap must indicate future resource requirements to avoid surprises with capacity and costs.

> *Roadmap: A continuously up to date cloud roadmap should be available for each business areas cloud initiatives and thoroughly analyzed prior to utilizing cloud services.*

2. Determine Application Profile: Literature suggests the need to size the application, map all dependencies and identify data repositories as well as integrations (Sabharwal & Wali 2013; Anderson 2018).

> *Application Profile: Understanding key characteristics of the application is important in order to build a more precise business case.*

3. Create & Analyze Business Case: Cloud sourcing should ensure business benefits are gained in addition to technical benefits of the cloud (Muhic & Bengtsson 2019; Willcocks et al. 2013). Furthermore, licensing costs are evident when migrating applications to the cloud. There are several different licensing options however, some scenarios are more beneficial than others, especially from a cost perspective. (Andrikopoulos et al. 2013; Mohan Murthy et al. 2013) The empirical results prove that it is important to evaluate the entire application stack when considering licenses, as well as understand that licensing varies case by case and should be kept in mind throughout the cloud journey.

> *Business Case Model: From a cost and capacity perspective, both sourcing and licensing are essential topics as they assist matching business justifications with the most cost-effective cloud options.*

4. Define & Evaluate Exit Plan: The empirical results suggest an exit plan should be defined regardless of costs. Literature similarly suggests that identifying when a cloud provider is not delivering services according to the agreement, finding a new cloud provider becomes relevant (Willcocks et al. 2013). The empirical results further suggest that exit plans are important to take into consideration especially when designing the application, and the applications cloud journey, to help prevent vendor lock-in.

5. Justify Business Case: Moving applications to the cloud must ensure benefits to the business. Cost benefits are pivotal when making a business case. (Preimesberger 2017 & Mithani et al. 2010). The empirical results also reveal the importance of cost considerations prior to moving applications to the cloud. Although costs might not be as important as the technical and business benefits of the cloud, they should be clearly indicated in the business case to ensure the cloud journey is well planned and in line with the case company's IT strategy.

6. Proceed to Cloud Plan: Once the business case has been approved and all the prior activities have been completed, the cloud journey can begin.

7. Subprocess 7.1-7.8: Cost Optimization & Capacity Management Process, Prior to Cloud, IaaS.

8. Subprocess 8.1-8.9: Cost Optimization & Capacity Management Process, Prior to Cloud, PaaS.

9. Review Architecture: A review of the architecture is necessary to ensure the correct methods and most cost optimal decisions have been made for the cloud journey. The empirical results prove the need to thoroughly examine all options, as down the line i.e. rehosting could be tweaked with revising. Literature similarly states how finding the most optimal deployment model for the application requirements and costs is difficult. This requires in-depth knowledge of the application and the chosen cloud environment. (Evangelinou et al. 2018; Tran et al. 2011). Moreover, Willcocks et al. (2013) suggest tapping into the innovation possibilities of the cloud for increased benefits of cloud computing. For these reasons, once phases 7 & 8 have been completed, more in-depth understanding of the application and cloud environment options should have been gained and therefore the architecture requires reviewal.

> *Service Model Evaluation: Before moving to the cloud, a final check of possible architectural changes must be conducted to identify improvement options.*
>
> *Continuous Improvement: The ability to evaluate service models and how they match different types of applications should be continuously improved to assist with upcoming cloud journeys.*

10. Proceed to Cloud: Once the architecture has been reviewed, the application can proceed to the cloud.

11. Subprocess 11.1-11.10: Continuous Cost Optimization & Capacity Management Process, In Cloud, IaaS.

12. Subprocess 12.1-12.13: Continuous Cost Optimization & Capacity Management Process, In Cloud, PaaS

13. Review Architecture: The empirical results suggest that it is good to review the cloud deployment and check for tradeoffs between i.e. agility and license costs and service models. Literature identifies how over time resources may shift to another type of workload as a result of changing business requirements, usage or costs (Microsoft Azure 2018). Empirical results and literature also highlight the importance of checking for innovation possibilities of the cloud (Willcocks et al. 2013). For these reasons, the architecture requires reviewal to determine whether a new service model is needed.

> *Service Model Evaluation: After being in the cloud for a while, it is good to review the tradeoffs between service models. Improvement options should be evaluated to further optimize costs as well as enhance business and technical requirements.*

14. Proceed to Exit: Once the decision has been made to no longer continue with the current cloud provider, the exit process may begin.

15. Execute Exit Plan: The exit plan must be put into effect.

6.3.1 Prior to Cloud, IaaS

7.1 Confirm User & Usage Amount from Business: The empirical results indicate how confirming the user and usage amounts are important and directly relate to resource requirements. Literature highlights how end users and end user behavior generate workloads that are processed using cloud resources (Jennings & Stadler 2015).

7.2 Analyze Workload Pattern (Complete Cycle): If the application is currently on-premise, the workload pattern of the application requires analysis. It is important to keep in mind that the entire workload cycle is analyzed, as the empirical results prove that applications may vary immensely when it comes to workload patterns. Literature also highlights how existing tools provide historical details which can be used to determine workload patterns (Allspaw & Kelariwal 2017; Reese 2009).

> *Existing Tools (Native Tools for On-Premise Components): Native on-premise tools are a good place to start in order to begin gathering information on the applications behavior in terms of workload.*

7.3 Analyze On-Premise Component Capacity: Similarly, to 7.2, 7.3 entails the use of existing on-premise native tools to identify component capacity, such as CPU and RAM.

> *Existing Tools (Native Tools for On-Premise Components) CPU, RAM…: Correspondingly to workload patterns, native on-premise tools can be used to identify*

component capacity details, which assist in building the application capacity requirements.

7.4 Estimate Capacity Requirements: If the application is currently on-premise, then the capacity requirements can be estimated according to the details gathered in 7.1, 7.2 and 7.3. However, if the application does not exist on-premise and is completely new, an estimation will need to be made without the help of on-premise native tools, and instead be based on knowledge gained in 7.1.

Vendor Best Practices: Used to assist with estimating capacity requirements.

7.5 Rightsize & Forecast Capacity: Empirical results suggest that checking the current on-premise capacity requires further analysis to re-check if the amount of resources could be reduced. Literature identifies how the lowest possible amount of capacity should be used, and rightsizing the environment is critical in order to save costs (Sabharwal & Wali 2013; Amazon 2018). Furthermore, capacity forecasts are needed to understand future capacity requirements. Literature proves that accurate knowledge on demands and the ability to forecast load are key when it comes to using resources in the cloud (Reese 2009; Hu et al. 2014).

7.6 Compare On-Premise vs. Cloud Costs: Empirical results clearly depict the importance of making cost calculations. Calculations should highlight the capacity and usage requirements as well as take future forecasts into consideration.

7.7 Compare Cost Models: The different cost models should be compared, as the applications workload patterns affect the choice. Analyzing the different cost models is important to ensure the best decision is made from a technical and economic standpoint. (Suleiman et al. 2012; Jennings & Stadler 2015) The empirical results demonstrate how the different options should be weighed against each other to find the most optimal solution.

Activities 7.5-7.7 should be continuously iterated. Understanding workload and capacity requirements and comparing them to the available cost models and cloud offerings is important in order to figure out the best solution. Literature identifies how cost optimization should be conducted as a continuous process, where cost optimization options are evaluated, prioritized, and implemented as well as constantly improved (Cristea 2017). For this reason, these activities should be completed several times especially when the application may suit various cost model options or if the application is completely new. It is also important to keep in mind that the case companywide tagging strategy is enforced as the importance of tagging is evident from both the empirical results and theoretical background.

Vendor Best Practices: Assist with making more accurate decisions.

Microsoft Excel, AWS/ Azure Online Calculator: Used to make relevant calculations.

AWS/ Azure Online Documentation: Used to gain understanding on cost models, cost calculations and technical details, i.e. to know which cost model (RI, Savings Plans, On-Demand) and technical options (autoscale, on/ off, alerts, scripts) to take into use according to the workload.

7.8 Complete Final Cost & Capacity Design: Literature suggests that all the essential pre-requisites should be thoroughly examined, and then documented (Mithani et al. 2010). Once the most suitable environment has been identified from a cost and capacity point of view, the application can move forward in the cloud journey.

Lessons Learned: Continuously collected to assist with future cloud journeys.

6.3.2 In Cloud, IaaS

Literature points out how understanding the applications behavior in a cloud environment prior to moving the application to the cloud can be difficult especially from a cost perspective (Evangelinou et al. 2018). Depending on how well the applications capacity and costs were optimized in the Prior to Cloud Process will define whether the application needs to go through and be analyzed in the First Complete Workload Cycle (11.1) or moved directly to the Continuous Optimization Cycle (11.6).

11.1 Monitor Cost & Capacity: Literature suggests that monitoring capacity is essential in order to understand the applications resource requirements (Sabharwal & Wali 2013; Anderson 2018). The empirical results similarly suggest that the application requires monitoring in order to understand how the application uses resources in the cloud.

11.2 Analyze Cost & Capacity: Literature suggests in order to make decisions related to capacity requirements, thorough analysis of the monitored application is needed (Sabharwal & Wali 2013; Anderson 2018). Once the first complete workload cycle of the application is complete, the workload patterns and resource requirements of the application should be reassessed by comparing the different available cost models to ensure economic benefits. Empirical results suggest that the applications characteristics are important to understand in order to reach the most cost optimal option. Once again, the different options should be financially weighed against each other.

Activities 11.1-11.2 should be continuously iterated until a clear idea of the workload cycle is available and an appropriate decision can be made regarding the capacity and

costs of the application. Other factors such as version upgrades etc. should be kept in mind as they may bring cost savings. This became evident from the empirical results.

Vendor/ Case Company Best Practices: Assist with making more accurate decisions.

Existing Tools, AWS/ Azure Native Tools: Used to monitor resource usage.

AWS/ Azure Online Documentation: Used to gain understanding on cost models, and technical details, i.e. to know which cost model (RI, Savings Plans, On-Demand) and technical options (autoscale, on/ off, alerts, scripts) to take into use.

Cost Visibility Tool: Used to monitor costs.

11.3 Propose Change: Empirical results suggest that the appropriate vendor or cloud service partner should suggest the cost optimization and capacity management related changes.

ServiceNow: The proposed change should be requested via ServiceNow to centralize and facilitate monitoring the progression of the requests.

11.4 Review Change (Worth it?): The proposed change ticket via ServiceNow should be assigned to the applications IT Architect for reviewal, and to the IT Service Manager for approval. Literature suggests that cost optimization initiatives require prioritizing the worthiness of the change at hand (Cristea 2017). Empirical results further notify that each change should be analyzed in order to determine whether the change is worth it.

Change Evaluation: Savings potential, time investment, resource requirements and technical risks to be analyzed alongside each proposed change.

11.5 Implement Change: If additional approval is not required, the approved change can be performed. If the change is large and affects multiple applications, then approval is needed from IT Management. Once approval has been given, the change can be implemented. If the change is either rejected by the IT Service Manager or IT Management, then the process moves back to activity 11.1.

11.6 Standard Monitoring & Change Process: Literature clearly highlights the importance of implementing cost optimization initiatives as a continuous practice (Cristea 2017). Empirical results similarly point out how the process should be continuous, as cost optimization is very important to the case company. Activities 11.1-11.4 should be continuously carried out according to the appropriate time period. Some applications may require more frequent monitoring than others, depending on the chosen cost model. It is also important to keep business demands in mind, as well as continuously monitor resources

in order to identify any resources that are not utilized in a manner that is efficient cost or capacity wise (Sabharwal & Wali 2013; Blair & Chandrasekaran 2019).

11.7 & 11.9. Enforce Cost Optimization: Literature suggests that appropriate reporting should be available to the suitable parties (Sabharwal & Wali 2013). Empirical results similarly suggest that visibility of costs is highly important to understand the situation and optimization needs.

> *Cost Reports: Cost reports should be readily available and easily accessible. Visibility will enforce the optimization of costs, especially when missed savings potentials are recognized.*

11.8 Enforce Unforeseen Change in Capacity Requirements: Additional business demands that affect capacity should be identified and aligned with the roadmap. These changes need to be enforced and implemented according to 11.6 (11.1-11.4).

11.10 Implement Change: Once the approval process has been completed, the change can be implemented and once again activity 11.6 (11.1-11.4) begins. The roadmap should be kept in mind and continuously checked to anticipate future demands.

> *Roadmap: A continuously up to date cloud roadmap should be available for each business areas cloud initiatives and analyzed when utilizing cloud services.*

6.3.3 Prior to Cloud, PaaS

Activities 8.1-8.4 follow activities 7.1-7.4.

8.5 Analyze Cloud Platform Services: Literature highlights how defining requirements will need to be made according to the offerings and identified cloud platform services (Mithani et al. 2010; Willcocks et al. 2013; Anderson 2018). The empirical results further prove that the PaaS model is more complex than the IaaS one, therefore this activity requires additional analysis.

Activities 8.6, 8.7 & 8.8 correspond to activities 7.5, 7.6 & 7.7.

Activities 8.6-8.8 should be continuously iterated in order to find the best solution.

> *Vendor Best Practices: Assist with making more accurate decisions.*
>
> *Microsoft Excel, AWS/ Azure Online Calculator: Used to make relevant calculations.*
>
> *AWS/ Azure Online Documentation: Used to gain understanding on cost models, cost calculations and technical details, i.e. to know which cost model (Code On*

Demand, Subscription) and technical options (autoscale, tiers, plans, alerts, scripts) to take into use.

Activity 8.9 corresponds to activity 7.8.

6.3.4 In Cloud, PaaS

Depending on how well the applications capacity and costs were optimized in the Prior to Cloud Process will define whether the application needs to go through the Minimum Viable Development activities (12.1), the First Complete Workload Cycle (12.4) or moved directly to the Continuous Optimization Cycle (12.9).

12.1 Evaluate & Select Cheapest Plan: The empirical results identify how starting with the cheapest plan is ideal when building a new development. This assists with understanding the way a new development uses resources and accumulates costs. Literature highlights how there are many options in the cloud therefore, the applications should be tested in a cloud environment. In addition, designing and development of applications should be done in a way that consumes the lowest possible amount of resources. (Sabharwal & Wali 2013; Hähnle & Johnsen 2015)

12.2 Assess Viability of Selected Plan: The selected plan should be thoroughly assessed to ensure the appropriate plan has been chosen.

12.3 Reassess Cost & Capacity Requirements: If costs and capacity are not according to plan, the situation requires reassessment.

Activities 12.1-12.3 should be continued until the most appropriate plan cost and capacity wise has been identified.

Existing Tools, AWS/ Azure Native Tools: Used to monitor and assess the developments.

AWS/ Azure Online Documentation: Used to gain understanding on cost models, cost calculations and technical details, i.e. to know which cost model (Code On Demand, Subscription) and technical options (autoscale, tiers, plans, alerts, scripts) to take into use.

12.4 Monitor Cost & Capacity: Once the plan is ok cost and capacity wise, the application can move to the First Complete Workload Cycle process. Empirical results suggest that monitoring assists to help identify when changes need to be made to i.e. the applications code to stabilize the application.

12.5 Analyze Cost & Capacity: The cost and capacity should be assessed to identify improvement areas. Empirical results suggest that the better and more thought out the

application design, the more cost optimized the result. Literature similarly highlights how the design is key in order to avoid unnecessary costs (Hähnle & Johnsen 2015).

Activities 12.4-12.5 should be continuously iterated until a clear idea of the workload cycle is available, and an appropriate decision can be made regarding the capacity and costs of the application. Other factors such as new technical capabilities, upgrades etc. should be kept in mind as they may bring cost savings. This became evident from the empirical results.

> *Vendor/ Case Company Best Practices: Assist with making more accurate decisions.*
>
> *Existing Tools, AWS/ Azure Native Tools: Used to monitor resource usage.*
>
> *AWS/ Azure Online Documentation: Used to gain understanding on cost models, and technical details, i.e. to know which cost model (Code On Demand, Subscription) and technical options (autoscale, tiers, plans, alerts, scripts) to take into use.*
>
> *Cost Visibility Tool: Used to monitor costs.*

Activities 12.6-12.8 are identical to 11.3-11.5. If the change is rejected, the process moves back to 12.4.

12.9 Standard Monitoring & Change Process: Similarly, to 11.6, activity 12.9 should be continuous.

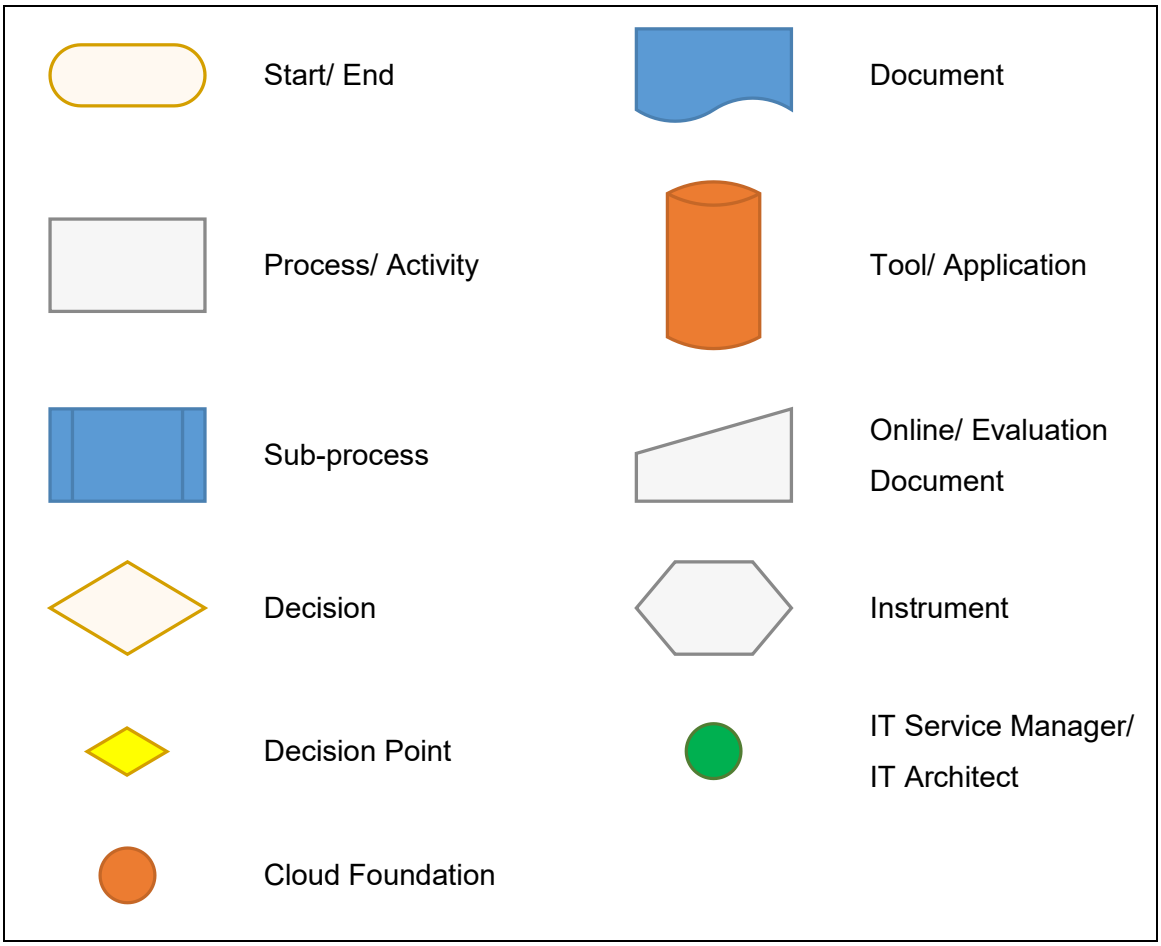
Activities 12.10-12.12 are identical to 11.7-11.9.

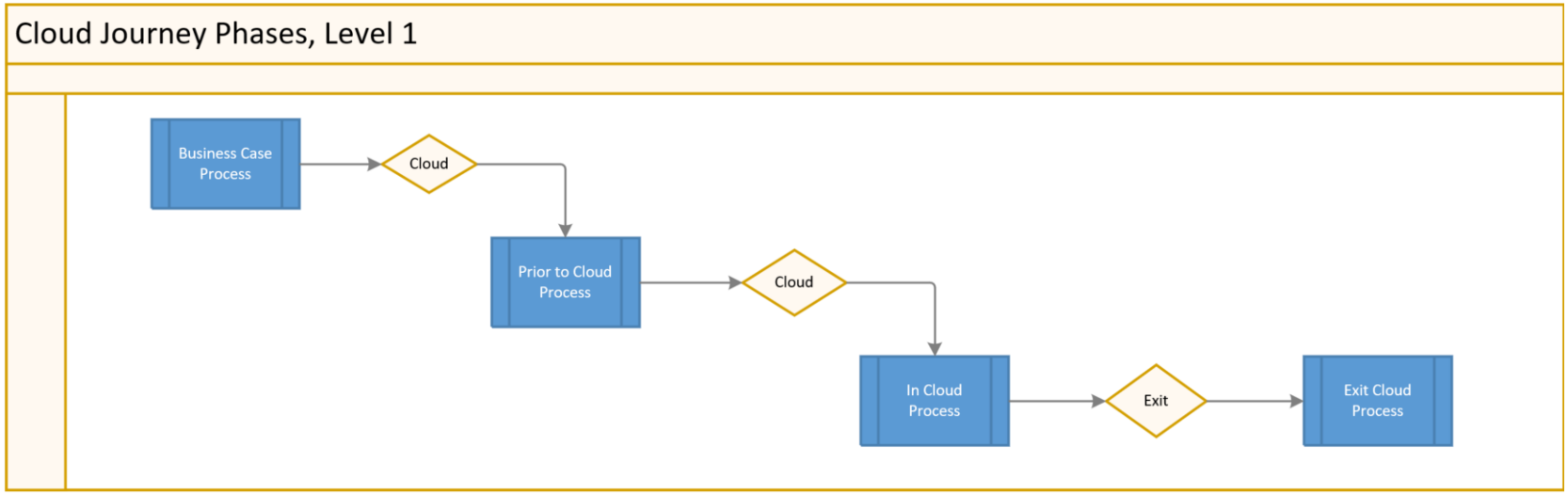
12.13 Implement Change: Once the approval process has been completed, the change can be implemented and once again activity 12.9 (12.4-12.7) begins. The roadmap should be kept in mind and continuously checked to anticipate future demands.

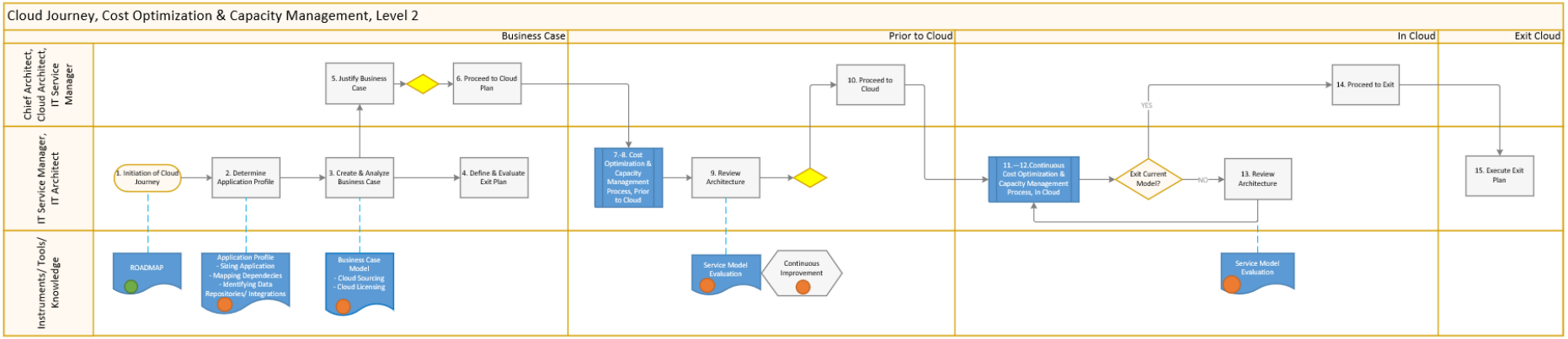
> *Roadmap: A continuously up to date cloud roadmap should be available for each business areas cloud initiatives and analyzed when utilizing cloud services.*

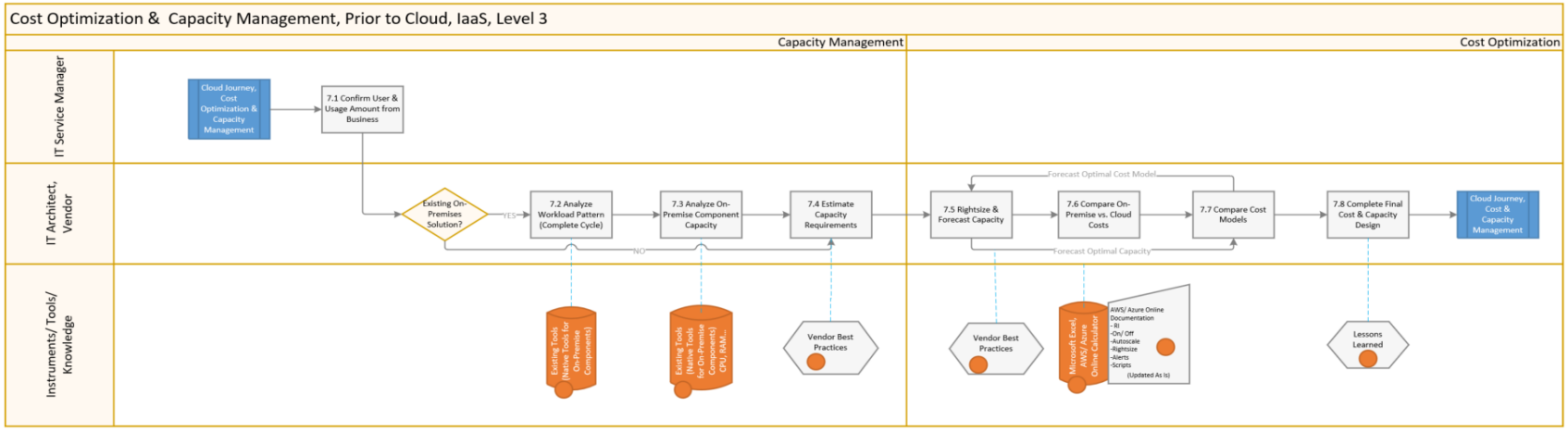
6.4 Process Drawings

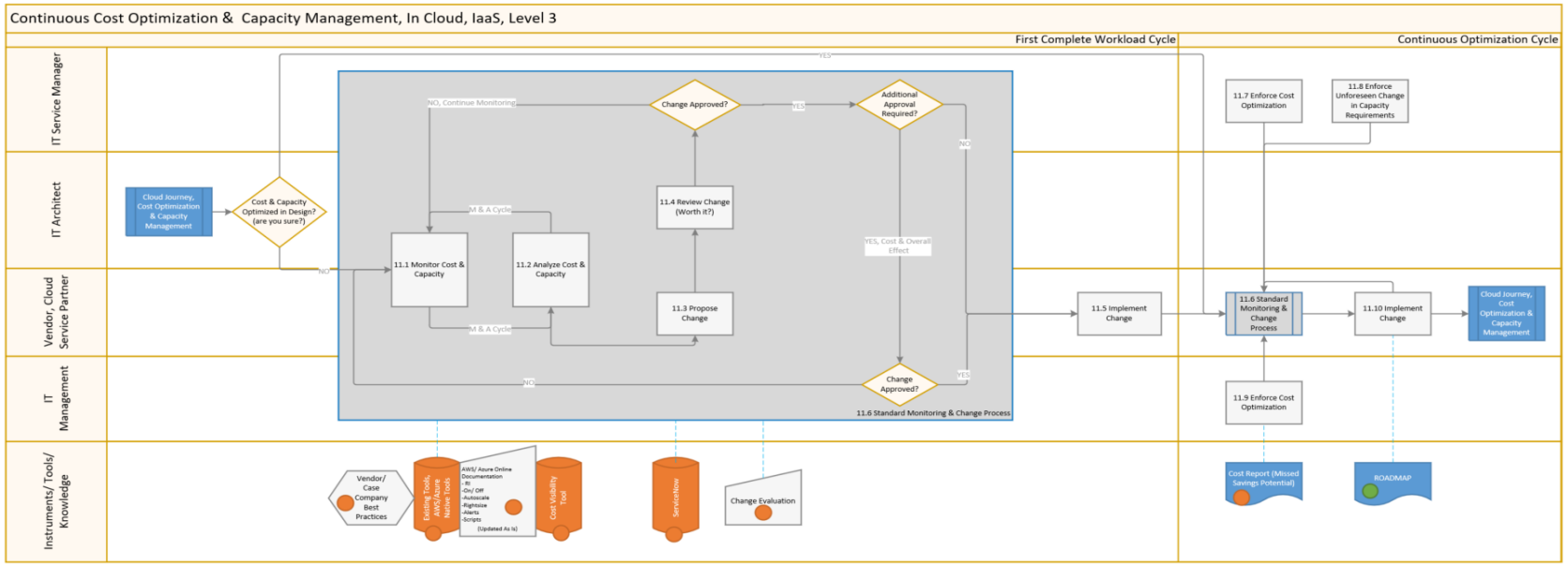
This section includes the process drawings which were detailed in sections 6.2 and 6.3. The symbols used throughout the process are presented prior to the processes. By zooming in, the processes will be more readable.

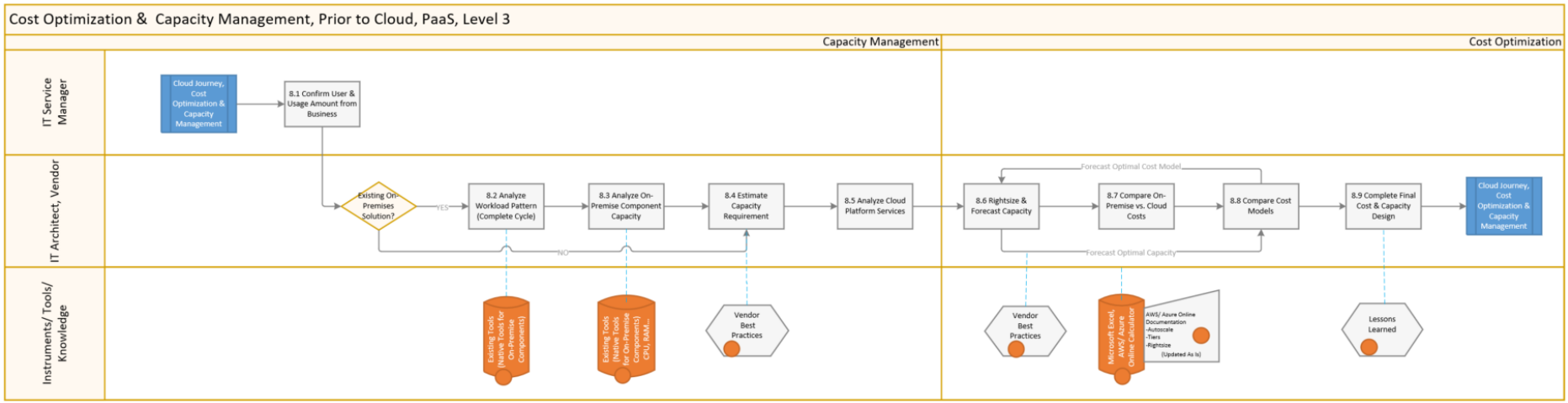

Start/ End
Document
Process/ Activity
Tool/ Application
Sub-process
Online/ Evaluation Document
Decision
Instrument
Decision Point
IT Service Manager/ IT Architect
Cloud Foundation

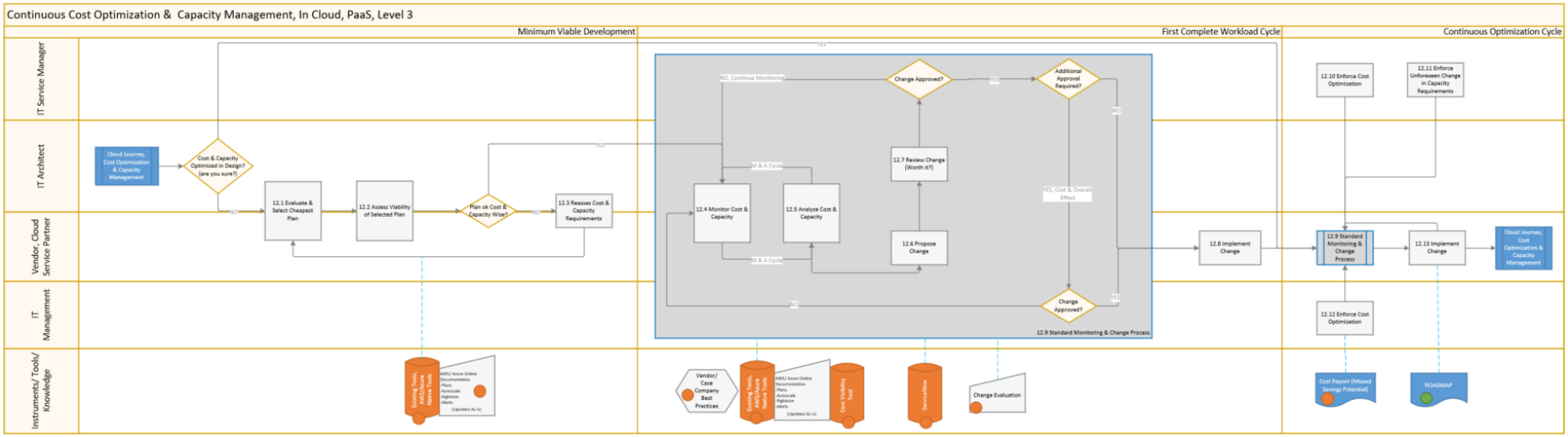

Cloud Journey Phases, Level 1
Business Case Process
Cloud
Prior to Cloud Process
Cloud
In Cloud Process
Exit
Exit Cloud Process


Cloud Journey, Cost Optimization & Capacity Management, Level 2
Business Case
Prior to Cloud
In Cloud
Exit Cloud
Chief Architect, Cloud Architect, IT Service Manager
IT Service Manager, IT Architect
Instruments/ Tools/ Knowledge
1. Initiation of Cloud Journey
2. Determine Application Profile
3. Create & Analyze Business Case
4. Define & Evaluate Exit Plan
5. Justify Business Case
6. Proceed to Cloud Plan
7.-8. Cost Optimization & Capacity Management Process, Prior to Cloud
9. Review Architecture
10. Proceed to Cloud
11. – 12. Continuous Cost Optimization & Capacity Management Process, In Cloud
Exit Current Model?
YES
NO
13. Review Architecture
14. Proceed to Exit
15. Execute Exit Plan
ROADMAP
Application Profile
- Sizing Application
- Mapping Dependencies
- Identifying Data Repositories/ Integrations
Business Case Model
- Cloud Sourcing
- Cloud Licensing
Service Model Evaluation
Continuous Improvement
Service Model Evaluation


Cost Optimization & Capacity Management, Prior to Cloud, IaaS, Level 3
Capacity Management
Cost Optimization
IT Service Manager
IT Architect, Vendor
Instruments/ Tools/ Knowledge
Cloud Journey, Cost Optimization & Capacity Management
7.1 Confirm User & Usage Amount from Business
Existing On-Premises Solution?
YES
NO
7.2 Analyze Workload Pattern (Complete Cycle)
7.3 Analyze On-Premise Component Capacity
7.4 Estimate Capacity Requirements
Forecast Optimal Cost Model
Forecast Optimal Capacity
7.5 Rightsize & Forecast Capacity
7.6 Compare On-Premise vs. Cloud Costs
7.7 Compare Cost Models
7.8 Complete Final Cost & Capacity Design
Cloud Journey, Cost & Capacity Management
Existing Tools (Native Tools for On-Premise Components)
Existing Tools (Native Tools for On-Premise Components) CPU, RAM...
Vendor Best Practices
Vendor Best Practices
Microsoft Excel, AWS/ Azure Online Calculator
AWS/ Azure Online Documentation
- RI
- On/ Off
- Autoscale
- Rightsize
- Alerts
- Scripts
(Updated As Is)
Lessons Learned


Continuous Cost Optimization & Capacity Management, In Cloud, IaaS, Level 3
First Complete Workload Cycle
Continuous Optimization Cycle
IT Service Manager
IT Architect
Vendor, Cloud Service Partner
IT Management
Instruments/ Tools/ Knowledge
Cloud Journey, Cost Optimization & Capacity Management
Cost & Capacity Optimized in Design? (are you sure?)
YES
NO
NO, Continue Monitoring
Change Approved?
YES
Additional Approval Required?
NO
M & A Cycle
11.4 Review Change (Worth it?)
11.1 Monitor Cost & Capacity
11.2 Analyze Cost & Capacity
11.3 Propose Change
YES, Cost & Overall Effect
Change Approved?
YES
NO
11.6 Standard Monitoring & Change Process
11.5 Implement Change
11.7 Enforce Cost Optimization
11.8 Enforce Unforeseen Change in Capacity Requirements
11.6 Standard Monitoring & Change Process
11.10 Implement Change
Cloud Journey, Cost Optimization & Capacity Management
11.9 Enforce Cost Optimization
Vendor/ Case Company Best Practices
Existing Tools, AWS/Azure Native Tools
AWS/ Azure Online Documentation
- RI
- On/ Off
- Autoscale
- Rightsize
- Alerts
- Scripts
(Updated As Is)
Cost Visibility Tool
ServiceNow
Change Evaluation
Cost Report (Missed Savings Potential)
ROADMAP


Cost Optimization & Capacity Management, Prior to Cloud, PaaS, Level 3
Capacity Management
Cost Optimization
IT Service Manager
IT Architect, Vendor
Instruments/ Tools/ Knowledge
Cloud Journey, Cost Optimization & Capacity Management
8.1 Confirm User & Usage Amount from Business
Existing On-Premises Solution?
YES
NO
8.2 Analyze Workload Pattern (Complete Cycle)
8.3 Analyze On-Premise Component Capacity
8.4 Estimate Capacity Requirement
8.5 Analyze Cloud Platform Services
Forecast Optimal Cost Model
8.6 Rightsize & Forecast Capacity
8.7 Compare On-Premise vs. Cloud Costs
8.8 Compare Cost Models
8.9 Complete Final Cost & Capacity Design
Forecast Optimal Capacity
Cloud Journey, Cost Optimization & Capacity Management
Existing Tools (Native Tools for On-Premise Components)
Existing Tools (Native Tools for On-Premise Components) CPU, RAM...
Vendor Best Practices
Vendor Best Practices
Microsoft Excel, AWS/ Azure Online Calculator
AWS/ Azure Online Documentation
-Autoscale
-Tiers
-Rightsize
(Updated As Is)
Lessons Learned


Continuous Cost Optimization & Capacity Management, In Cloud, PaaS, Level 3
Minimum Viable Development
First Complete Workload Cycle
Continuous Optimization Cycle
IT Service Manager
IT Architect
Vendor, Cloud Service Partner
IT Management
Instruments/ Tools/ Knowledge
Cloud Journey, Cost Optimization & Capacity Management
Cost & Capacity Optimized in Design? (are you sure?)
NO
YES
12.1 Evaluate & Select Cheapest Plan
12.2 Assess Viability of Selected Plan
Plan ok Cost & Capacity Wise?
12.3 Reasses Cost & Capacity Requirements
NO, Continue Monitoring
Change Approved?
YES
Additional Approval Required?
NO
M & A Cycle
12.4 Monitor Cost & Capacity
12.5 Analyze Cost & Capacity
12.6 Propose Change
12.7 Review Change (Worth it?)
YES, Cost & Overall Effect
Change Approved?
NO
YES
12.9 Standard Monitoring & Change Process
12.8 Implement Change
12.10 Enforce Cost Optimization
12.11 Enforce Unforeseen Change in Capacity Requirements
12.9 Standard Monitoring & Change Process
12.13 Implement Change
Cloud Journey, Cost Optimization & Capacity Management
12.12 Enforce Cost Optimization
Vendor/ Case Company Best Practices
Existing Tools, AWS/Azure Native Tools
AWS/ Azure Online Documentation
-Plans
-Autoscale
-Rightsize
-Alerts
(Updated As Is)
Cost Visibility Tools
ServiceNow
Change Evaluation
Cost Report (Missed Savings Potential!)
ROADMAP

6.5 Recommendations

The process, which was based on the findings made from the empirical results and the theoretical background and modeled according to the PKM framework, proved the need to implement a KM process along the process itself. When modeling the process, it became evident that majority of the instruments, knowledge and tools of the KM process belong to the cloud foundation, as it is currently the only centralized cloud service within the case company. The case company has many internal and external parties therefore, a centralized group of individuals should overlook the entire process and ensure it is kept up to date and implemented.

The cloud foundation must establish practices for the key instruments which include, lessons learned, continuous improvement and best practices. These are required to ensure cost optimization and capacity management activities progress and become more advanced and accurate. In addition, the cloud foundation should have the ability to guide individuals on how to use and find the tools and knowledge identified along the process. Furthermore, the cloud foundation should create clear requirements for the different knowledge documents of the process.

To ease the flow of the process, ServiceNow needs to be used as a centralized tool. ServiceNow should have up to date processes, data and knowledge, which enable the appropriate parties to perform the activities along the process. In addition, ideally the cost reports would be integrated into ServiceNow from the native AWS and Azure tools, with appropriate access rights to the data. Dashboard availability should be constant. Cost reports should allow the end user to view costs from a single application level, to case companywide business area level cost reports. Centralizing the prior mentioned factors into one tool simplifies the overall process landscape.

One factor which became highly evident throughout the interviews was the importance of application specific knowledge. The empirical results state how a centralized service for optimization would be ideal however, application specific knowledge and details typically reside with individuals who work closely with the application. For this reason, individuals specialized with the application need to be kept in the loop of any ongoing cost optimization and capacity management activities. The proposed cost optimization and capacity management changes should be primarily made by the vendor or cloud service partner. The approvals on the other hand, need to come from the case company. However, when the resource mass grows large enough cost optimization and capacity management activities should be centrally controlled.

Furthermore, as evident from the empirical results, the technical offerings of public clouds are constantly evolving. Keeping up with the offerings and benefits as well as matching them to potential cloud applications is extremely difficult for individuals who are not as familiar with the cloud. For this reason, the cloud foundation should have the ability to centrally assist and provide consultation for application cloud journeys that need support.

7. CONCLUSIONS

In this chapter the conclusions are presented. Chapter 7.1 presents the research questions and how they were answered. Chapters 7.2 and 7.3 discuss theoretical and practical contributions. Chapter 7.4 discusses limitations and chapter 7.5 presents suggestions for future research.

7.1 Meeting the Objectives of the Research Questions

How can effective cloud cost optimization and capacity management support the optimization of cloud costs?

Literature and the empirical results highlight the importance of cost optimization as a design. In addition, existing literature clearly depicts how cost optimization needs to be considered as an ongoing practice. Therefore, the process created as a result of this thesis takes cost optimization into consideration throughout the different phases of the cloud journey, emphasizing on the importance of design in the planning phase, and creating an ongoing process for the run phase activities.

Capacity management is pivotal when considering costs in a cloud environment. Both literature and empirical results prove how the chosen capacity design directly correlates with the accumulated costs. As capacity management and cost optimization go hand in hand, the different capacity related options require constant evaluation and consideration from a cost perspective. Therefore, in order to achieve cost optimization, capacity management must be conducted. Capacity management is taken into consideration by pointing out essential capacity related activities throughout the process, which includes the planning and the run phase.

How to design business processes to account for cost optimization and capacity management?

The PKM framework was used to help build the process. The framework entailed identifying processes, activities, instruments, tools, knowledge, roles and responsibilities. By combining the literature review and empirical results, cost optimization and capacity management processes could be identified. The activities along the process, as well as the instruments, tools, knowledge, roles and responsibilities were all combined into one process.

7.2 Theoretical Contribution

Existing literature identifies how the use of public cloud services has increased in recent years. The increasing use of cloud services has led to a rise in costs, which organizations from varying industries are currently facing. Cost optimization and capacity management have been identified as important factors that affect costs. Therefore, this thesis explores the connection between cost optimization and capacity management. The theoretical contribution of this thesis is a process that takes cost optimization and capacity management into account.

The process created as a result of this thesis combines knowledge gained from the theoretical background and the empirical results with the help of the PKM framework. Literature often introduces capacity management processes from the point of view of the cloud provider (Sabharwal & Wali 2013). Practice based models on cost optimization exist from a cloud consumers point of view (Anderson 2018). In addition, the complexity of the cloud as well as the amount of possible solutions and the differences in costs are discussed in literature (Koziolek et al. 2011; Evangelinou et al. 2018) and are evident from the empirical results. This thesis introduces a cost optimization and capacity management process from a cloud consumers point of view. Furthermore, this thesis demonstrates a KM process that supports cost optimization and capacity management activities along the business process while taking the rapidly evolving cloud environment into account.

The theoretical background and empirical results jointly demonstrate the relationship between cost optimization and capacity management and the importance of a process that connects both aspects. Empirical results prove how cost optimization and capacity management choices, such as choosing a cost model according to capacity directly relates to costs. Activities along the business process and the instruments, knowledge and tools of the KM process in chapter 6 are discussed in literature however, they are often separate entities. For instance, articles may focus on the technical aspects of capacity management, or on cost optimization topics such as service models, cost models or licensing. As cost optimization and capacity management impact costs, a process which takes both factors into account is necessary in order to assist in controlling the issue with rising costs. This thesis further combines different cost optimization and capacity management topics from the literature review and empirical results, into one business process. In addition, the KM process supports the activities along the business process, by further detailing relevant instruments, knowledge and tools for cost optimization and capacity management.

The most central theoretical contribution of this thesis is a process which combines cost optimization and capacity management from existing literature and the empirical results. The process focuses on this topic from a cloud consumers perspective, which is currently not commonly available in literature.

7.3 Practical Contribution

The practical contribution of this thesis is a process which combines cost optimization and capacity management for the planning and the run phase activities of the cloud journey. The process has been created from a cloud consumer point of view. Therefore, organizations planning to move applications to a cloud environment can identify pivotal cost optimization and capacity management related activities from the process in order to minimize and tackle cost related issues.

The results of thesis can be generalized, as the process primarily created for the case company can be used by other organizations that have a similar set-up and are approximately the same size. Cloud consumers should focus on the entire cloud journey process and especially emphasize the importance of the Prior to Cloud phase. However, organizations must also ensure the implementation of an In Cloud phase. In order to achieve this, organizations must understand the benefits of the KM process, which includes central instruments, knowledge and tools, that support the entire cloud journey process.

7.4 Limitations

The result of this thesis is a process that is based on the buildup of different cost optimization and capacity management activities identified from the literature review and empirical study. Therefore, the overall benefit of the process would need to be tested in practice as well as further investigated and studied in order to analyze the actual savings gained from the identified cost optimization and capacity management activities.

The lack of previous academic research on the topic affects the reliability of the research in this thesis. Current research on the topic of cloud cost optimization is evident in practice-based models. However, academic research mainly focuses on cost optimization and capacity management as separate entities, and no clear process that combines both factors is available. In addition, continuous advancements in cloud computing technology are evident, which causes knowledge to become outdated in a rapid manner.

In addition, the sample used for the empirical study consists of employees from a single organization. Therefore, the empirical results reflect the inputs of solely one organization.

Furthermore, the samples were all from the IT department of the organization. The samples included different roles within the IT department. Although the varying roles of the samples within the case company gave a wide overlook on the topic, it slightly limited the possibility to examine the topic at a more specific level. However, this enabled the process to include a variety of aspects, ranging from more technical, to management level viewpoints.

The cloud journeys in this thesis consist of three different phases, the planning, migrating and in cloud phases. The samples were at different parts of the three phases. Therefore, the different phases limit the reliability of the overall results, as several samples were not in the cloud environment yet. Moreover, the different service models of the samples further limit the reliability. In other words, the varying phases and service models cause individual views to become overemphasized, limiting the results of this thesis.

7.5 Suggestions for Future Research

As the use of cloud computing continues to grow in organizations, future research should increasingly explore cloud computing from the perspective of cloud consumer organizations. During the empirical study it became evident that the design of an application plays an extremely important role in the cloud journey, and how the different service and cost models bring varying benefits to consumers. Future research should study how the choices made at the beginning of a cloud journey effect the tradeoffs between benefits related to business, technology and costs. Short-term and long-term tradeoffs should be compared between the different service and cost models, to determine the most prominent issues and benefits that arise from the decisions made at the beginning.

In addition, vendor lock-in should be further explored. The risk of vendor lock-in varies depending on the chosen service and cost model. The importance of an exit plan was evident from the empirical results. For this reason, future research should examine how vendor lock-in can be avoided efficiently. Moreover, the tradeoffs between the risk of vendor lock-in and the effort placed into avoiding vendor lock-in should be analyzed.

Future research should also explore the available tools on the market that assist with cost optimization and capacity management when moving from an on-premise to a cloud environment. The ability to understand the change in capacity between on-premise and cloud environments, as well as the different service and cost models could be simplified with the use of a tool. Therefore, future research should identify the benefits of this type of tool, and especially the tradeoffs between the cost of the tool, and the potential benefits the tool can bring cost optimization and capacity management wise.

REFERENCES

Alkhalil, A., Sahandi, R., & John, D. (2017). An exploration of the determinants for decision to migrate existing resources to cloud computing using an integrated TOE-DOI model. Journal of Cloud Computing, Vol. 6(1), pp. 1-20.

Allspaw, J. & Kejariwal, A. (2017). The Art of Capacity Planning, 2nd Edition. O'Reilly Media.

Allweyer, T. (1999). A Framework for Redesigning and Managing Knowledge Processes. Saarbrücken.

Amazon. (2018). Cost Optimization Pillar. AWS Well-Architected Framework, Available: https://d1.awsstatic.com/whitepapers/architecture/AWS-Cost-Optimization-Pillar.pdf

Amazon. (2019). Amazon EC2 Pricing. Available: https://aws.amazon.com/ec2/pricing/

Amazon. (2019b). AWS Auto Scaling. Available: https://aws.amazon.com/autoscaling/

Amazon. (2019c). AWS Instance Scheduler. Available: https://aws.amazon.com/solutions/instance-scheduler/

Anderson, E. (2018). Cost Optimization Using Cloud Computing. Gartner.

Andrikopoulos, V., Binz, T., Leymann, F., & Strauch, S. (2013). How to adapt applications for the Cloud environment. Computing, Vol. 95(6), pp. 493–535.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., & Zaharia, M. (2009). Above the clouds: A Berkeley view of cloud computing. University of California, Berkeley.

Barroso, L., A. & Hölzle, U. (2007). The Case for Energy-Proportional Computing. Computer, Vol. 40(12), pp. 33–37.

Blair, R. & Chandrasekaran, A. (2019). 10 Best Practices for Azure Cloud IaaS Cost Optimization. Gartner, ID: G00343592.

Block, D. (2012). Governing the cloud as cloud-based services evolve, so must today's governance functions. KPMG. pp. 1-4.

Cancila, M. (2015). How to Budget, Track and Reduce Public Cloud Spending. Gartner, ID: G00272868.

Case Company. (2019a). Case Company website.

Case Company. (2019b). Enterprise Architecture Repository.

Chaisiri, S., Lee, B-S., & Niyato, D. (2009). Optimal virtual machine placement across multiple cloud providers. 2009 IEEE Asia-Pacific Services Computing Conference (APSCC), pp. 103-110.

Chang, V., Walters, R., & Wills, G. (2013). The development that leads to the Cloud Computing Business Framework. International Journal of Information Management, Vol. 33(3), pp. 524–538.

Clayton, T. (2018). Decision Point for Choosing a Cloud Migration Strategy for Applications. Gartner, ID: G00361356.

Cristea, A. M. (2017). Cost Optimization as Managerial Strategy in the Context of Increasing the Complexity of Inter-Functional Decision-Making Process. Hyperion International Journal of Econophysics & New Economy, Vol. 10(2), pp. 189–199.

De Capitani Di Vimercati, S., Foresti, G., Livraga, V., Piuri and P., Samarati, (2013) Supporting User Requirements and Preferences in Cloud Plan Selection. IEEE Transactions on Services Computing.

Debreceny, R.S. (2013). Research on IT governance, risk, and value: challenges and opportunities. Journal of Information Systems, Vol. 27(1). pp. 129-135.

Evangelinou, A., Ciavotta, M., Ardagna, D., Kopanel, A., Kousiouris, G., Varvarigou, T. (2018). Enterprise applications cloud rightsizing through a joint benchmarking and optimization approach. Future Generation Computer Systems, Vol. 78(1), pp. 102-114.

Ganly, C. & Naegle, R. (2019). Driving Cost Optimization Across the Enterprise: An IT Perspective. Gartner, ID: G00383464.

Gomolski, B., Kost, J. (2009). Decision Framework for Prioritizing Cost Optimization Ideas. Gartner, ID: G00166206.

Han, Y. (2011). Cloud Computing: Case Studies and Total Costs of Ownership. Information Technology & Libraries, Vol. 30(4), pp. 198–206.

Hayes, J. (2010). Clouding the licensing issues? [enterprise software licensing]. Engineering and Technology, Vol. 5(17), pp. 52–53.

Huang, D., Yi, L., Song, F., Yang, D., & Zhang, H. (2014). A secure cost-effective migration of enterprise applications to the cloud. International Journal of Communication Systems, Vol. 27(12), pp. 3996–4013.

Hu, R., Jiang, J., Liu, G., & Wang, L. (2014). Efficient Resources Provisioning Based on Load Forecasting in Cloud. The Scientific World Journal.

Hähnle, R. & Johnsen, E.B. (2015). Designing Resource-Aware Cloud Applications. Computer, Vol. 48(6), pp. 72-75.

Jennings, B., Stadler, R. (2015). Resource Management in Clouds: Survey and Research Challenges. Journal of Network & Systems Management, Vol. 23(3), 567–619 (2015).

Jiang, Y., Perng, C., Li, T. & Chang, R. (2012). Intelligent cloud capacity management. 2012 IEEE Network Operations and Management Symposium, pp. 502-505.

Kavis, M. (2014). Architecting the cloud: Design decisions for cloud computing service models (SaaS, PaaS, and SaaS). Hoboken, New Jersey: Wiley.

Khoury, G.R. (2010), Innovative Cost Optimization. A Creative Approach to Finding New Cost Optimisation Opportunities. Available: http://gkstrategic.com/pdf_image/Innovative%20Cost%20Optimisation%20-%20Gerald%20Khoury15.pdf

Koziolek, A., Koziolek, H., Reussner, R. (2011). PerOpteryx: Automated application of tactics in multi-objective software architecture optimization. Proceedings of the Joint ACM SIGSOFT Conference -- QoSA and ACM SIGSOFT Symposium -- ISARCS on Quality of Software Architectures -- Qosa and Architecting Critical Systems – Isarcs. pp. 33–42.

KPMG. (2008). Cost optimization, protecting our margins in a turbulent economic environment.

Lněnička, M. (2013). Cloud Based Testing of Business Applications and Web Services. Scientific Papers of the University of Pardubice. Series D, Faculty of Economics & Administration, Vol. 18(26), pp. 66–78.

Loten, A. (2018). Rush to the Cloud Creates Risk of Overspending. The Wall Street Journal. Available: https://blogs.wsj.com/cio/2018/07/25/rush-to-the-cloud-creates-risk-of-overspending/

Louridas, P. (2010). Up in the air: Moving your applications to the cloud. IEEE Software, Vol. 27(4), pp. 6-11.

Lubrecht, M.D., Pizzo, K.A., Savvides, A., Baron, A., & Papaefstathiou, E. (2010). Methods for capacity management.

Maier, R. (2002), Knowledge management systems. Information and communication technologies for knowledge management, Springer, Berlin.

Maier, R. & Remus, U. (2003) Implementing process-oriented knowledge management strategies. Journal of Knowledge Management, Vol. 7(4), pp. 62-74.

Malik, T., Chard, K., & Foster, I. (2014). Benchmarking cloud-based tagging services. 2014 IEEE 30th International Conference on Data Engineering Workshops, pp. 231-238.

Maresova, P., Sobeslav, V., & Krejcar, O. (2017). Cost–benefit analysis – evaluation model of cloud computing deployment for use in companies. Applied Economics, Vol. 49(6), pp. 521–533.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., Ghalsasi, A. (2011). Cloud computing — The business perspective. Decision Support Systems, Vol 51(1), pp. 176-189.

Martens, B., Walterbusch, M. & Teuteberg, F. (2012). Costing of cloud computing services: A total cost of ownership approach. Proceedings of the 2012 45th Hawaii International Conference on System Science (HICSS'12). IEEE, pp. 1563-1572.

Matthews, B., Jones, C., Puzo, B., Moon, J., Tudhope, D., Golub, K., & Lykke Nielsen, M. (2010). An evaluation of enhancing social tagging with a knowledge organization system. Aslib Proceedings, Vol. 62(4), pp. 447-465.

Mell, P., & Grance, T. (2010). The NIST Definition of Cloud Computing. Communications of the ACM, Vol. 53(6), pp. 50.

Microsoft Azure. (2018). Best practices for costing and sizing workloads migrated to Azure. Available: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/migrate/azure-best-practices/migrate-best-practices-costs

Microsoft Azure. (2019). Build a cost-conscious organization. Available: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/organize/cost-conscious-organization

Microsoft Azure. (2019b). Track costs across business units, environments, or projects. Available: https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/ready/azure-best-practices/track-costs

Mithani, M. F., Salsburg, M. A., & Rao, S. (2010). A Decision Support System for Moving Workloads to Public Clouds. International Journal on Computing, Vol. 1(1), pp. 150-157.

Mohan Murthy, M., Ameen, M., Sanjay, H., & Yasser, P. (2013). Software Licensing Models and Benefits in Cloud Environment: A Survey. Advances in Intelligent Systems and Computing, Vol. 174, pp. 645-650.

Muhic, M., Bengtsson, L. (2019). Dynamic capabilities triggered by cloud sourcing: a stage-based model of business model innovation. Review of Managerial Science.

Muhic, M., Johansson, B. (2014). Cloud Sourcing – Next Generation Outsourcing? Procedia Technology – Elsevier, Vol. 16(C), pp. 553-561.

Nissen, M., Kamel, M., & Sengupta, K. (2000). Integrated analysis and design of knowledge systems and processes. Information Resources Management Journal, pp. 24-43.

Ojala, A. (2013). Software-as-a-Service Revenue Models. IT Professional, Vol. 15(3), pp. 54–59.

Peiris, C., Balachandran, B. & Sharma, D. (2010). Governance framework for cloud computing. International Journal on Computing, Vol. 1(1). pp. 88-93.

Prasad, A., Green, P. (2015). Governing cloud computing services: Reconsideration of IT governance structures. International Journal of Accounting Information Systems, Vol 19, pp. 45-58.

Prasad, A., Green, P. & Heales, J. (2014). On governance structures for the cloud computing services and assessing their effectiveness. International Journal of Accounting Information Systems, Vol 15(4), pp. 335-356.

Preimesberger, C. (2017). 10 Mistakes to Avoid When Migrating Data Centers to the Cloud. EWeek, pp. 1.

Reese, G. (2009) Cloud application architectures. O'Reilly Media, Inc., Sebastopol.

Roseline, T., Tauro, C. & Miranda, M. (2017). An approach for efficient capacity management in a cloud. 2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), pp. 1-6.

Rountree, D. & Castrillo, I. (2014). The Basics of Cloud Computing: Understanding the Fundamentals of Cloud Computing in Theory and Practice. Syngress.

Sabharwal, N. & Wali, P. (2013). Cloud Capacity Management. Apress.

Schneider, S., & Sunyaev, A. (2016). Determinant factors of cloud-sourcing decisions: Reflecting on the IT outsourcing literature in the era of cloud computing. Journal of Information Technology, Vol. 31(1), pp. 1-31.

Singh, S. & Chana, I. (2015). Q-aware: Quality of Service based cloud resource provisioning. Computers & Electrical Engineering, Vol. 47, pp. 138-160.

Singh, S. & Chana, I. (2015). QRSF: QoS-aware resource scheduling framework in cloud computing. The Journal of Supercomputing, Vol. 71(1). pp. 241-292.

Suleiman B., Sakr S., Jeffery R., Liu A. (2012). On understanding the economics and elasticity challenges of deploying business applications on public cloud infrastructure. Journal of Internet Services and Applications, Vol 3, pp. 173-193.

Sultan, N., & van de Bunt-Kokhuis, S. (2012). Organisational culture and cloud computing: coping with a disruptive innovation. Technology Analysis & Strategic Management, Vol. 24(2), pp. 167–179.

Sumalatha, K., & Anbarasi, M. S. (2019). A review on various optimization techniques of resource provisioning in cloud computing. International Journal of Electrical & Computer Engineering, Vol. 9(1), pp. 629–634.

Tak, B. C., Urgaonkar, B., & Sivasubramaniam, A. (2013). Cloudy with a Chance of Cost Savings. IEEE Transactions on Parallel & Distributed Systems, Vol. 24(6), pp. 1223-1233.

Teece, D.J. (2018) Business models and dynamic capabilities. Long Range Planning, Vol. 51(1), pp. 40–49.

Tran, V., Keung, J., Liu, A., & Fekete, A. (2011). Application migration to a cloud: a taxonomy of critical factors. SECLOUD '11 Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing.

Vaquero L., Rodero-Merino L., & Buyya, R. (2011) Dynamically scaling applications in the Cloud. ACM SIGCOMM Computer Communication Review, Vol. 41(1), pp. 45–52.

Vithayathil, J. (2018) Will cloud computing make the information technology (IT) department obsolete? Information Systems Journal, Vol 28(4), pp. 634–649.

Wang, Z., Hayat, M.M., Ghani, N., & Shaban, K., B. (2017). Optimizing Cloud-Service Performance: Efficient Resource Provisioning via Optimal Workload Allocation. IEEE Transactions on Parallel and Distributed Systems, Vol. 28(6), pp. 1689-1702.

Ward, J., & Slattery, T. (2018). The rise of cloud computing. Accountancy Ireland, Vol. 50(1), pp. 22-23.

Weinman, J. (2012). Cloudonomics: The Business Value of Cloud Computing. John Wiley & Sons. Hoboken. pp. 160.

Wiebe, E., Durepos, G., & Mills, A. J. (2010). Encyclopedia of Case Study Research. Los Angeles [Calif.]: SAGE Publications, Inc.

Willcocks L.P., Venters W., Whitley E.A. (2013). Cloud sourcing and innovation: slow train coming? A composite research study. Strategic Outsourcing: An International Journal, Vol. 6(2), pp. 184–202.

Wu, C., Buyya, R., & Ramamohanarao, K. (2019). Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges. ACM Computing Surveys, Vol. 52(6), pp. 1-36.

APPENDIX A: CLOUD JOURNEY PLANNING PHASE INTERVIEW TEMPLATE

Introduction & Business Justification

1. Could you briefly explain what your role is in the planning phase of the applications cloud journey?
2. Which parties are/ will need to be present in the planning and run phase activities?
3. What is the business justification and are there any expected business outcomes or objectives of the applications cloud journey?

Prior to the Cloud

4. How important is cost optimization prior to moving the application to the cloud?
5. How is the amount of capacity needed estimated before moving the application to the cloud?
6. Which tools will be used for capacity and spend related forecasts?
7. Who will be in charge of estimating the capacity needs and creating the capacity design?
8. What types of problems have come up so far with the capacity need estimations?

In the Cloud

9. How important is cost optimization in the run phase?
10. How will the capacity be monitored? (Tools)
11. Who will be in charge of managing and following the capacity related details?
12. Who will apply the changes?
13. How often will capacity management related activities take place?
14. When will cost optimization be worth it?

General

15. Have you thought about licenses from a cost optimization perspective?
16. Have you thought about any exit plan if costs begin to rise?
17. What have been the most important lessons learned so far?

18. Do you have any expectations on how the case company should conduct cost optimization as a service? (Centralized service, what type of service, what type of data etc. would you like to see)

APPENDIX B: CLOUD JOURNEY MIGRATION/ IN CLOUD PHASE INTERVIEW TEMPLATE

Introduction & Business Justification

1. Could you briefly explain what your role was/ is in the planning and run phase of the applications cloud journey?
2. Which parties were/ are present in the planning phase and are currently a part of the run phase activities?
3. What was the business justification and were/ are there any expected business outcomes or objectives of the applications cloud journey?

Prior to the Cloud

4. How important was cost optimization prior to moving the application to the cloud?
5. How was the amount of capacity needed estimated before moving the application to the cloud?
6. Were any tools used for capacity and spend related forecasts?
7. Who was in charge of estimating the capacity needs and creating the capacity design?
8. What types of problems came up with capacity need estimations?

In the Cloud (the below questions were slightly reworded for interviewees in the migration phase):

9. How important is cost optimization in the run phase?
10. How is the applications capacity managed in the cloud?
11. How is the capacity monitored? (Tools)
12. Who is in charge of managing and following the capacity related details?
13. Who applies the changes?
14. How often do capacity management related activities take place?
15. When is cost optimization worth it?
16. Are there any (reoccurring) problems with capacity management in the run phase?

General

17. Have you thought about licenses from a cost optimization perspective?
18. Is there any exit plan if costs begin to rise?
19. How accurate were the forecasts capacity and spend wise to the actual reality of the deployment in the cloud?
20. What were the most important lessons learned during the cloud journey?
21. Do you have any expectations on how the case company should conduct cost optimization as a service? (Centralized service, what type of service, what type of data etc. would you like to see)