Tragedy of the Common Cloud

Game Theory on the Infrastructure-as-a-Service Market

Jörn Künsemöller

Dissertation zur Erlangung des Grads *Doktor der Naturwissenschaften*

Fakultät für Elektrotechnik, Informatik und
Mathematik der Universität Paderborn

13. Mai 2014

1. Gutachter: Prof. Dr. Holger Karl
2. Gutachter: Prof. Dr. Claus-Jochen Haake

"I do not fear computers.
I fear the lack of them."

– *Isaac Asimov*

Abstract

Recent Internet technology advancements enable and simplify the outsourcing of corporate and private information technology to remote facilities. Organizing this as public services has become known as *cloud computing*. This thesis investigates the prospective market development for cloud storage and processing services (*Infrastructure-as-a-Service*). The main focus is to identify important factors with a huge influence on market form and pricing, which should hence be considered in the choice and standardization of services and in market regulation.

Different aspects of the cloud market are modeled and analyzed by means of game theory. Pricing in a monopoly market is explored in regard to a possible combination of public cloud services with an own infrastructure (*hybrid cloud*). For competitive markets, the existence of stable market situations and how they are affected by complex tariff structures is investigated. Further, the separation of processing and storage facilities in different locations can provide different locational advantages for service providers and their users.

Competing only in price turns out to be without much potential for a sustainably stable market. In a monopoly, the possibility to build hybrid clouds appears to be essential for a relatively small on-demand service price. Differences between providers in their facility location and production costs have potential for stable market shares and prices. Legal frameworks and financial interests can support storage-only data centers that also might be distributed over the network in order to take part in future Internet technologies.

Zusammenfassung

Der Ausbau des Internets ermöglicht und vereinfacht zunehmend die Auslagerung informationstechnischer Aufgaben aus Unternehmen und Privathaushalten heraus an spezialisierte Dienstleister. Ist dies in Form von öffentlich verfügbaren Diensten organisiert, spricht man von *Cloud Computing*. Diese Arbeit befasst sich mit dem Markt für Rechen- und Speicher-Dienste der Cloud (*Infrastructure-as-a-Service*), dessen zukünftige Entwicklung untersucht wird. Im Fokus steht dabei die Identifizierung von Faktoren, die großen Einfluss auf Marktform und Preisbildung haben und daher bei der Standardisierung oder Regulierung des Marktes berücksichtigt werden sollten.

Verschiedene Marktaspekte werden mit spieltheoretischen Methoden modelliert und analysiert. Unter Berücksichtigung einer möglichen Kombination öffentlicher Dienste mit eigener Infrastruktur (*Hybrid Clouds*) wird die Preisbildung in einem Monopol untersucht. Es wird geprüft inwiefern sich aus verschiedenen komplexen Preismodellen stabile Marktsituationen im Anbieter-Wettbewerb ergeben. Außerdem wird die räumliche Trennung von Rechen- und Speicherinfrastruktur und die Streuung der Gesamtkapazität auf verschiedene Standorte erforscht, aus der sich verschiedene Standortvorteile für Anbieter und Nutzer ergeben können.

Es zeigt sich, dass der Wettbewerb über den Preis allein wenig Potenzial für dauerhafte Stabilität liefert. Im Monopol erweist sich die Möglichkeit der Nutzung von öffentlichen Diensten in Hybrid Clouds als wesentliche Voraussetzung für vergleichsweise niedrige Nutzungsgebühren. Unterschiede bezüglich Standort oder Produktionskosten ermöglichen stabile Marktanteile mehrerer Wettbewerber. Rechtliche Rahmenbedingungen und ökonomische Interessen bieten das Potenzial von reinen Speicher-Datenzentren, die im Rahmen neuer Internet-Technologien gegebenenfalls über die Netztopologie verteilt werden sollten.

Contents

"There are no rules of architecture
for a castle in the clouds"

– Gilbert K. Chesterton

1 Introduction

1.1 Motivation

More and more individuals and businesses seek to utilize the Internet in order to optimize their *Information Technology (IT)*. Under the premise that clustering similar tasks results in their more efficient accomplishment, information processing and storage tasks may be better outsourced to a remote provider that serves several clients. Such a provider might also be located in a more appropriate location than client IT facilities.

The outsourcing of IT tasks to the Internet can be organized as paid public services. This has become known as *cloud computing* (Section 2.1). An appropriate service payment creates a win-win situation to provider and clients: the cost-effectiveness of the provider allows a profitable service price that is below the client's production costs of this service. The difference of price and production costs is often considerable and can comprise orders of magnitude in particular cases [23].

There are important factors other than money, of course. Little capital expenditure and high flexibility favor outsourcing, but other factors do not. The protection of sensible information from unauthorized access is an issue when such data is processed or stored in a public service and legal regulations have to be considered in this context. Next to security, availability and fault liability of a service might argue against cloud computing as well. Some might not

even consider the cloud due to these reasons while other ventures, especially small start-ups, would not even be possible without it. A large number of potential clients, though, can be expected to balance pros and cons of cloud services against an own IT infrastructure and costs are especially important to those clients with large processing and storage demand.

Cloud services are already used in substantial amount today. Gartner estimates *Infrastructure-as-a-Service (IaaS)* revenue at about $6.2 billion in 2012 [39]. This is a relatively small amount compared to the global data center hardware investments which are estimated at $106.4 billion [40]. Because of the huge possible savings to clients, though, it is quite unlikely that cloud infrastructure utilization will remain a rare case in our competitive economic system. Cloud computing has the potential to completely rearrange IT infrastructure.

1.2 Working Hypothesis & Approach

This thesis studies the potential development of the IaaS market. Economies of scale create a cost advantage that makes cloud services favorable over in-house services and large providers over smaller ones. This promotes a redeployment of corporate IT as a whole in form of a monopolistic cloud provider. The working hypothesis is the existence of key factors and conditions that have a huge influence on this process and the prospective market form. Such aspects have to be taken into account when forming common cloud standards. Knowledge of these key factors and their influence is important for an educated and reflected utilization of cloud services and – if applicable – market regulation.

The market development is determined by the interaction of different actors: service providers and clients. There is need for a theoretical approach that can provide solutions for such situations. Game theory meets this requirement and is successfully used in manifold settings. Different market aspects are explored throughout this thesis using game-theoretic models. They shall provide answers to the following questions: Is a mass-movement of IT infrastructure to the cloud realistic? Is it of mutual benefit to all involved parties? How can this development be influenced, e.g. for market regulation?

1.3 Thesis Structure

Basic information about cloud computing and game theory is provided in Chapter 2. This defines the object of research of this thesis and gives the necessary methodological background in order to follow the subsequent chapters. The research findings from cloud market modeling are presented in Chapters 3 to 6, where each chapter focuses on a different market aspect. The dynamics of cloud instance price and utilization in a monopoly is investigated in Chapter 3, while Chapter 4 explores how this is affected by provider competition. Whether processing and storage facilities in separate locations can be competitive against facilities that provide both resources is analyzed in Chapter 5. While cloud computing is about to change the provision of IT resources, the Internet as the enabling medium is about to change as well. New needs require Internet architecture changes (often referred to as the *future Internet*) and key technologies that are up for debate often require some form of caching. Chapter 6 explores how this might influence the development of IaaS. A comprehensive discussion of the findings in Chapter 7 concludes this thesis.

1.4 Contributions

The following list outlines the main contributions of the individual research chapters and related earlier publications.

Chapter 3: Pricing & Usage Dynamics The dynamics of on-demand pricing and service usage are investigated in a two-stage game model for a monopoly IaaS market. The possibility of hybrid clouds (clouds plus own infrastructure) turns out to be essential in order that not only the provider but also the clients have significant benefits from on-demand services. Even if the client meets all demand in the public cloud, the threat of building a hybrid cloud keeps the instance price low. This is not the case when reserved instances are offered as well. Parameters like load profiles and economies of scale have a huge effect on likely future pricing and on a cost-optimal split-up of client demand between either a client's own data center and a public cloud service or between reserved and on-demand cloud instances. Parts of this chapter have been published previously in [60] and [62].

Chapter 4: Provider Competition This chapter investigates how cloud provider competition influences instance pricing in an IaaS market. When reserved instance pricing includes an on-demand payment in addition to a reservation fee (two-part tariffs), several providers might offer different price combinations one of which might be preferable to a client according to its load profile. We investigate a duopoly of providers and analyze stable market prices in two-part tariffs. Further, we study offers that allow a specified amount of usage free of charge (three-part tariffs). For symmetric providers, neither two-part nor three-part tariffs produce an equilibrium market outcome other than a service pricing that equals production cost, i.e. complex price structures do not signifi-

cantly affect the results from ordinary Bertrand competition. When the providers have different production costs, there is an infinite number of equilibria with two-part tariffs where usually both providers make positive profit. Three-part tariffs may increase the provider's equilibrium profits. This is collaborative work together with Sonja Brangewitz and has been published previously in [59].

Chapter 5: Data Centers for Processing and Storage in Separate Locations When processing and storage are obtained as Internet services, the actual location of the providing facility is undetermined. This chapter contributes a market model for separate processing and storage facilities in comparison to a combined approach. It can be shown that stable market constellations with separate service specific facilities are possible when certain conditions (market share, economies of scale and location) are met. Large parts of this chapter have been published previously in [61].

Chapter 6: Cloud Infrastructure and the Future Internet This chapter evaluates a new business model where ISPs charge content providers (CPs) for a caching service since they benefit from content distribution. Although ISP caching is potentially not in equilibrium, it turns out to be Pareto optimal at the right pricing, which encourages cooperation between content providers and ISPs. Cloud storage providers have an incentive to choose cache friendly physical locations for their facilities in order to provide the necessary storage capacity to the ISPs. Further, we show that ISP caching as a paid service can be in equilibrium when future benefits are considered and when the ISP neutralizes caching-related improvements of service quality for clients that do not pay for caching. This is joint work together with Nan Zhang, João Soares and Kimmo Berg that has been published previously in similar form in [64] and [63].

"There is Nothing so Theoretical as a Good Method."

– Anthony G. Greenwald

2 Fundamentals

2.1 Cloud Computing

The idea to outsource IT to the Internet is not a new one and has been practiced in web applications for decades [18]. Nevertheless, the term *cloud computing* became a popular new expression for such outsourcing over the last years. It was coined by Ramnath Chellappa as a notion of an economy-driven IT-paradigm [20]. What exactly cloud computing means is not yet settled, though. Research with the intention to find a consensus in prevalent definitions found not a single aspect in agreement [102]. Hence, the definition of cloud computing remains as vague as the symbolic depiction of arbitrary constituted networks, where the term presumably originates from (Figure 2.1).

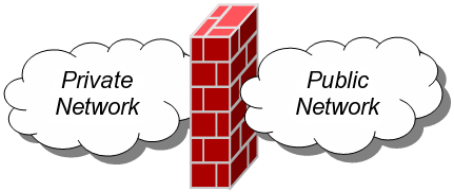


Figure 2.1: The cloud – a popular symbol for networks of arbitrary constitution; here for instance in the depiction of a firewall.

Throughout the following sections, an intentionally broad definition of cloud computing is described. It is meant to show where the marketing buzzword ends and cloud computing as an expression for a novel approach in

computing is actually justified. This definition determines the object of research for this thesis as far as required. It further gives an insight into the scope of more restrictive future definitions.

2.1.1 Service-Oriented Architecture

In order to grasp the basic idea of cloud IT, it is worthwhile to take a look at an earlier concept that has often been referred to in terms of web services: *Service-Oriented Architecture (SOA)*. This architectural paradigm, introduced by Gartner in 1996 [91], describes the approach to encapsulate processing tasks in discrete services. These services feature a defined interface that allows their utilization at appointed terms and conditions. The actual implementation and runtime environment remains hidden from the service user. Several services can be orchestrated to more complex applications. Such an application architecture allows systems based on different computer architecture or operating system to work in the same application. It also enables the reuse of implemented tasks in different contexts.

SOA is not a working technology with defined services and interfaces. It is an abstract concept that sets a flexible, modular design against a traditional monolithic structure. Web services can be regarded as a category for Internet-accessible IT-services of any kind following the architectural paradigm of SOA.

2.1.2 Cloud-Services and Economics

As mentioned in the previous section, the actual implementation of a service remains invisible behind its interface. An application that makes use of a service only has to know the interface. The actual implementation of the service does not have to be available until it is actually used. This means that the ser-

vice provider (responsible for the required implementation and resources in order to comply with the specified service) can remain undetermined when the application is built and can be chosen and varied at runtime. Cloud computing adds an economic component to this concept by introducing service fees. The charged amount grows with the amount of service usage. So first of all, cloud computing is a restructuring of sales and distribution: sell access to a product instead of the product itself. This creates a market economy where different implementations of the same service can compete in price and quality.

Although several services with a usage-based charging model exist, the market is still very heterogenous. There is no defined pattern at present into which new providers can easily integrate. Instead, there are lots of different provider-specific standards for e.g. processing and storage instances. The cloud market is still missing uniform standards for services of the same kind and an accepted set of services with fields of duties for each kind. These would make the presence of several providers for the same service a lot more likely and give an incentive for an (e.g. automatic) choice amongst competing providers at runtime.

Apparently, there also is no accepted open marketplace where providers and clients can bargain over cloud services, yet. But there is the perspective of such a market-driven SOA implementation of IT services on the Internet. And it is this prospect that makes cloud computing an expression for Internet outsourcing not only of former unknown quantity but also quality: a "computing paradigm where the boundaries of computing will be determined by economic rationale rather than technical limits alone." [20]. When considering that cloud services cover all areas of IT today, from application logic to hardware, cloud computing finally realizes a 50 year old idea:

> "computation may someday be organized as a public utility"
>
> – *John McCarthy, 1961* [28]

2.1.3 The Cloud Stack

Protocol stacks are often used to segment computer systems into defined fields of functions. These stack patterns consist of several layers of different degree of abstraction. The higher a layer is in the stack, the more general it is. Each layer covers one of those fields. The well-known *Open Systems Interconnection (OSI)* reference model, for instance, defines seven layers in data transfer over networks: from physical transmission to application-level tasks like file transfer [108]. Network protocols should cover one of these scopes. Such a modular approach offers huge variety like many different network technologies for instance. At the same time, unnecessary redundancy on other levels is avoided.

As long as the interfaces between the stack's layers are complied with, a layer can be implemented independently from the other layers. When developers follow such a pattern and integrate their product in a stack, implementations of the more general layer can make use of it. Growing acceptance of a stack increases the potential user base. Implementations for competing stack models are incompatible and become less usable, the more established the incompatible stack becomes. It hence becomes more promising to integrate new implementations into the most established stack. This works against a permanent heterogeneity throughout the evolution of a system. While this focuses implementation efforts, it unfortunately also makes it difficult to introduce a different structure.

With protocol stacks in mind, a similar framework for cloud computing is not far-fetched. Many contributions to a definition of cloud computing introduce such a cloud stack model. Most of these presentations suggest to present a predominant model, but the different models are actually quite varied in comparison (Figure 2.2).

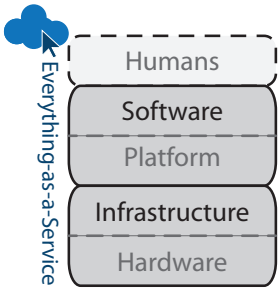


Figure 2.2: The cloud stack – a consistent model is not established.

Usually, a stack model of three layers is discussed, similar to the service models in the popular cloud computing definition by the *National Institute for Standards and Technology (NIST)* [73]. It differentiates tasks between hardware and end-user application in three scopes: *Software-as-a-Service (SaaS)*, *Platform-as-a-Service (PaaS)* and IaaS. According to this model, application logic is executed by a runtime environment based on (virtualized) hardware.

Other models suggest a larger number of layers. They add an additional layer for even more abstract, human-solvable tasks to the stack or sub-divide existing layers [15]. Infrastructure services, for instance, can be defined in a layer that abstracts virtual hardware from a layer of physical hardware. Occasionally, a less complex stack model is presented, where the PaaS-layer is omitted [10]. Platforms – software that is not geared to end users but provides a foundation for end-user applications, e.g. in form of libraries – are just regarded as SaaS in this simplified model. They can be used by other software services, if feasible, but there is no separate layer from software that targets end users. What remains is the traditional distinction of software and hardware, which kind of calls the idea of a stack model into question. Nevertheless, a provider-independent, uniform interface between these layers would be desirable.

Which stack model will prevail is hard to tell and the outcome of market evolution. A feasible model cannot simply be developed at the drawing board because proprietary solutions by large providers have a huge impact on what application ranges and interfaces become common for services. Instead, it is likely that the market agrees on one or a number of successful common standards. Projects like openstack [79] and eucalyptus [33] already adopt successful standards and make them available for self-service and to providers other than the one that introduced the standard in the first place.

Considering the broad spectrum of software and IT in general, a certain variety is quite likely to persist, despite of what was said above. Special *Service Level Agreement (SLA)*-requirements could promote the evolution of several incompatible cloud stacks, where each stack model targets a certain computation milieu. Specialized products for certain areas are already pooled in categories like *High-Performance-Computing-as-a-Service (HPCaaS)*. These products might be too unattractive in different context to give incentive to integrate them in a common standard. Something similar happened to the above mentioned OSI model, which was not adopted in many areas in favor of the four-layer model TCP/IP. It is also not settled whether it makes sense to combine completely different infrastructure types like processing and storage in the same layer. Specialized interfaces are quite likely in this area and different hierarchical structures are also possible.

Cloud computing developed a typical naming scheme, as can be observed in Figure 2.2 and in the above example of HPCaaS. This convention of naming services is called *Everything-as-a-Service (XaaS)* and distinguishes cloud services from other implementations of the same type (that are not accessible via web services) by annexing *-as-a-Service* after the type's name.

The research in this thesis focuses on cloud services for processing and storage. Such services require physical hardware and are primary IT resources of general purpose. Note that despite the use of the term IaaS, no specific

stack model definition is assumed. Certain demarcation characteristics (like a separation of processing and storage) are rather investigated in the following chapters.

2.2 Game Theory

> "In a game, each one tries to be smarter than the others. Game theory investigates the outcome of everybody trying to do so. And it treats the whole world as if it were a huge game."
>
> – *Christian Rieck* (translated from german) [88]

With these few words, the german economist Christian Rieck states the basic idea of game theory. The quote characterizes a domain that was started in the 19th century in economics by Augustin Cournot [25], evolved to a general theory until the middle of the 20th century especially by John von Neumann's work [76] and got adopted and further developed in many areas since then. Because there are many books on the topic that introduce game theory at great length, there is no need for another detailed portrayal. Nevertheless, the following sections provide an insight into concepts and terms of non-cooperative game theory. The descriptions are not exhaustive and more specific concepts are described later wherever they are used. This chapter is of appropriate brevity and meant to provide the necessary background to follow this thesis without further research or qualification and relies on the expertise of [87] where not indicated otherwise.

2.2.1 Basic Approach

The expression *game* usually names an activity where several players try to reach a defined objective. This happens in a competitive way in most games.

The game's rules determine admitted actions and the sequence of moves. Game theory formalizes games with the desire to understand decision making by the players. Contrary to decision theory, scenarios in game theory are multidimensional: a player's decision is not only relevant for him- or herself, but also for other players. This means that the environment of each player is affected by the decisions of other players. When making a decision for a move, this diverse situation has to be factored in. This can be done by anticipating other player's moves. In a game with a chronological order of the moves, decisions are taken relating to a situation that is produced by the preceding move. Hence, there is an indirect relation of a player's move and the resulting game situation. The choice of an action is strategic, as it is intended to cause others to move in a helpful way.

> "Game theory accordingly is a theory of social interaction"
>
> – *Christian Rieck* (translated from german) [87]

Intending to understand real situations instead of board or card games, the situations in question are formalized as a game. For this purpose, several factors of the situation are mapped into a set of rules: actors, their possible actions, sequence of moves, and so on. Some outcomes of the game are more desirable to a player than others. This is represented by a *utility* value which is determined for each player and each possible outcome of the game. The higher the utility of a situation to a player, the more desirable it is to him or her.

Several forms of game depiction are common, each one particularly suitable for games of a certain complexity. Simple games are usually represented as n-dimensional matrices that contrast the possible moves of each player with one another. Each cell of the matrix represents a combination of all players' moves, which is a possible game result. The different utilities such a situation

offers to the players are stated in the cell. This is usually done in the form of a vector. In this thesis, cells are instead split diagonally. This makes it easier to comprehend which value belongs to which player.

The famous example of the *prisoners' dilemma* [37] is given in Figure 2.3 in order to clarify the formalization of a situation and its representation in *normal form*. The game models the case of two culprits who only can be fully convicted for their crime if at least one of them makes a confession. Both players move at the same time and can either CONFESS or DENY. If both deny, they are discharged. If only one confesses, he or she is rewarded as a key witness and the other one is punished. If both confess, both receive an eased punishment due to their cooperation. The cases of punishment are represented as negative numbers, acquittal is zero and the reward is positive. It is also perfectly possible to just make use of positive values as long as a more desirable outcome to the same player offers a higher utility value to him or her. In this example, both players have the same possible actions and utilities. A game like this is called *symmetric*. Nevertheless, if required by the situation, games where players have different utilities or possible actions can be constructed in the same fashion.

Figure 2.3: *Prisoners' dilemma* in strategic normal form.

Solutions & Solution Concepts

Game theory provides concepts to solve game models. A player's *strategy* is a specification of all decisions that the player will take throughout the game. A

game's *solution* is a combination of each player's strategy. The progression and outcome of the game is completely determined by such a strategy combination. *Solution concepts* systematically reduce the set of all possible courses of the game to a subset of solutions. Each solution is a suggestions on how a game is best played.

Each player is regarded a *homo oeconomicus* who intends to maximize his or her own utility. Because all players pursue this objective, a solution does not necessarily include the strategy that belongs to the outcome that is of maximum utility to a specific player. If the same situation is of low utility to another player, he or she will anticipate the move and – if possible – choose a strategy that avoids this situation. In anticipation of this other player's behavior, a player better pursues an outcome of suboptimal utility which can be realistically reached.

A solution may turn out to be a strategy combination that results in a lower utility to all players than what is offered by another game result. The prisoners' dilemma, for example, offers a relatively good utility to both players in case of an acquittal. This requires both culprits to deny. The reward gives an incentive to confess if the other one denies. A confession would also ease the penalty that a confession of the other culprit would cause. Hence, a confession is always the better option than denial, no matter what the other one does, Confess *dominates* Deny. Eliminating dominated strategies is maybe the most simple solution concept in game theory. If applied on the prisoners' dilemma, this leaves only one solution: both confess.

Players usually select the move that offers the biggest utility at a given move by the others. A strategy that is not dominated by another one might or might not be the best move, depending on what the other players do. In most games, there are certain strategy combinations where the chosen strategy of one player is the best response to the chosen strategies of all other players: there is no incentive for a unilateral change. Such a strategy combination is

in *Nash equilibrium* [74]. The identification of Nash equilibria is the essential solution concept next to dominance. The concept can be exemplified by the – quite stereotyping – example *battle of the sexes* [71]. It models a situation of a couple's evening plans. The situation is that they have a date, but neither of them remembers whether they decided on football or opera. Without the ability to communicate, each one has to simply decide for one location. As can be seen in Figure 2.4, the man has a higher utility from meeting the woman at the football match compared to meeting her at the opera and vice versa. If they happen to go to separate locations, this provides no utility to anybody. There are no dominant strategies in the game, but two Nash equilibria (in pure strategies): one where both choose FOOTBALL and one where both choose OPERA.

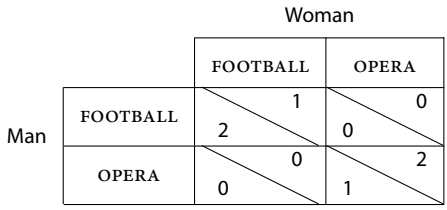


Figure 2.4: The *battle of the sexes* features two Nash equilibria.

Depending on the game, solution concepts like the Nash equilibrium may produce a varying number of solutions. Most other existing solution concepts are *refinements* of the Nash equilibrium that produce a more or less reduced subset of Nash equilibria. By selecting a reasonable concept, one can attempt to identify a single solution of the game that e.g. postulates how real players would actually play the game. Within this thesis, normal form representations and the Nash equilibrium concept are widely used e.g. in Chapters 4 and 6.

Mixed and Continuous Strategies

In addition to the described solutions in pure strategies, where the moves of each player are determined, there are solutions with so-called *mixed* strategies. A mixed strategy defines a variety of possible moves each of which is played with a specified probability [77]. Not every game has a Nash equilibrium in pure strategies, but all games feature an equilibrium in mixed strategies. It is defined by the mixed strategy of each player that offers a constant utility to other players (independent of their move). In the well-known game *rock, scissors, paper*, for instance, there is a mixed Nash equilibrium when all players follow the same strategy and play any possible move with a probability of $\frac{1}{3}$. In an asymmetrical game, though, the mixed strategies in equilibrium are usually not the same for each player. For instance, the *battle of sexes* (Figure 2.4) features a third equilibrium in mixed strategies (next to the two equilibria in pure strategies). While the man's probability for football is $\frac{2}{3}$, the woman goes there only with a probability of $\frac{1}{3}$.

It is sometimes hard to map all possible actions into the game model. A price, for example, can be determined in virtually unlimited increments and accordingly cannot be modeled by a finite number of pure strategies. This can be covered by *continuous* strategies, where a player chooses the value of a decision variable (instead amongst a finite number of explicitly modeled discrete values). Continuous strategies are used in Chapters 3 and 4 to cover instance price options.

Chronology

As long as possible actions at any given moment of the game can be determined and the utilities of possible game results stay unaltered throughout the game, game theory can model a situation with a chronological sequence of actions. The *one-shot* games presented above involve only one *simultaneous* move of each player. If such a game is repeated, each player can change its

decision between iterations and react on other players' moves. Strategies for repeated games, though, may be best given in form of abstract instructions like *tit for tat*. Other circumstances may not include a recurring situation. Sometimes, players have to move in specific order and possible moves may change because of earlier decisions. A player's strategy defines all moves of this player at any moment (and any possible former moves by other players) of the game. While all possible strategies could be represented in normal form by a row or column each, it usually is better readable in *extensive form* [76], a tree representation of the game.

The *battle of the sexes* (Figure 2.4) implies that no agreement can be made during the game. Because both move simultaneously, the move of the other person cannot be observed, either. If the game is modified so that one moves after the other, the one who moves first (and this move is observed by the other one) can factor in the behavior of the other one. Depending on who moves first, there are two different solutions of the game (Figure 2.5). If the woman begins, she will prefer OPERA over FOOTBALL, because the result of the expected man's move is of higher utility to her. If the man moves first, he can choose FOOTBALL without hesitation, knowing that the woman will follow. Both solutions can be identified by the Nash equilibrium concept. The example shows that the move order can be very important and has to be modeled in order to identify the right solution.

If the game would be continued in a way where both players take turns to redecide on their location, nobody would have an incentive to leave the place (since both would loose their utility if the other stays). This is one of the reasons why Nash equilibria are regarded to be good solutions: they render a game static since nobody can improve its situation on their own terms. This is indicated at the bottom of Figure 2.5. Chapter 3 uses an extensive form game to model a monopoly market for cloud infrastructure.

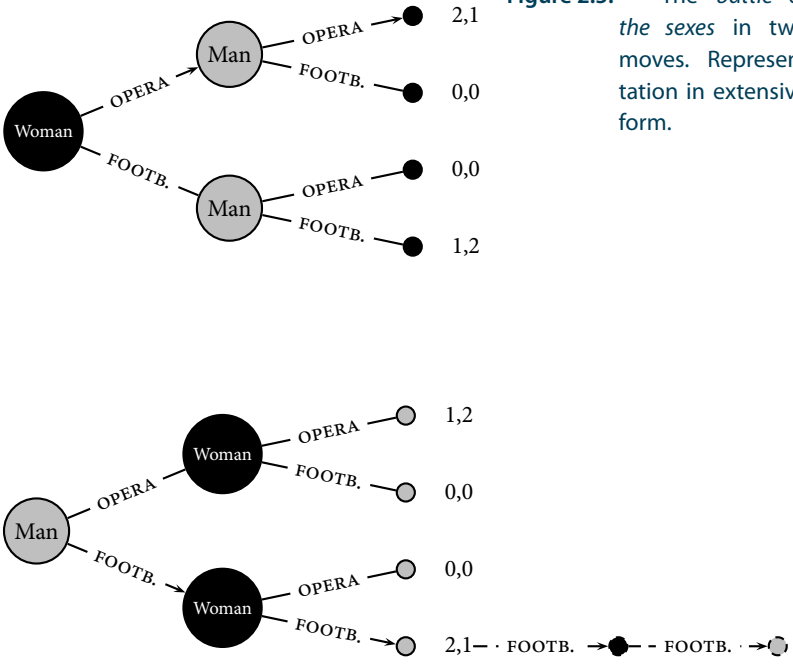


Figure 2.5: The *battle of the sexes* in two moves. Representation in extensive form.

2.2.2 Applicability & Limits

The plot of many standard models may not sound very credible. The *battle of the sexes* exemplifies an oversimplification of a complex social situation in an unrealistic game model. The purpose of these models is not to be a good representation of a real-world situation, but to offer a simple example for a standard situation in game-theoretic models. Thus, the situation that is supposedly modeled should rather be regarded a *cover story* in order to make the example more comprehensible and should not be taken too seriously. Stories like the *prisoners' dilemma* have become expressions for the game situation

they illustrate. The models in Section 2.2.1 lack many conditions of the real situation. In order to produce proper results, it is important to factor everything of significant influence on the actual decision into the model. Otherwise, utility assessments may be faulty or possible moves are not represented in the game. Note that the concepts of dominance and Nash equilibrium (in pure strategies) depend only on situation preference and hence abstract from discrete values. It is relevant how utilities are related to each other, the amount of the utility difference is not. Furthermore, only utility relations for the same player matter. Hence, a game model can produce plausible results as long as unaccounted influences do not change the order of precedence of the game results (Figure 2.6).

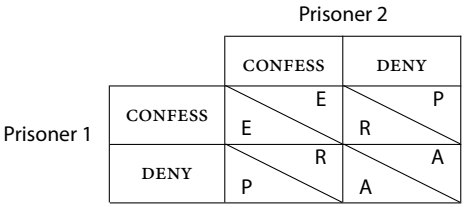


Figure 2.6: The *prisoners' dilemma* condition: P < E < A < R
(P)enalty, (E)ased penalty, (A)cquittal, (R)eward

Rationality

When possible actions change over time, for instance in a chess game, things soon get quite complicated. When a player decides on a move, he or she has to anticipate what choices become available to the other player by each possible own move. The player has to anticipate further what situations may emerge from each possible move of the other, and so on. Ideally, a player has to anticipate all possible game developments until the end in order to fully evaluate a strategy. Practically, this is very challenging. While a computer might be able to calculate a tree of possible game developments over a large number of moves, human players soon reach the limits of either their capabilities or

motivation. In repeated games, a player's strategy can be based on experience instead of rational thoughts, though.

Game theory requires *rationality* in a players choice of strategy in order to ensure that decisions are comprehensible. Rationality is taken for granted when players strive for maximal utility and e.g. do not randomly lower their utility without an apparent reason. How much utility a situation offers to a player, though, depends on the individual player's goals and values. How these can be assessed is the field of research of *utility theory*. In an economic context, money is an important utility measure of course.

Evaluating the utility of a situation only by its current pros and cons can be insufficient. The end result of a game might be the key factor in many games, but sometimes intermediary situations are important as well. This especially holds true for games that are played over significant time. A high average salary usually is of higher utility than a top amount right before retirement. Hence, there can be meta-level reasons for choosing a certain strategy that one would call irrational. These have to be factored into the utilities in order to consider them in the model. This can be done, for example, by summing up all utilities of preceding intermediary situations into the end result. If such a reasonable utility calculation cannot be arranged, game theory reaches its limits.

In addition, the use of *tactics*, meaning a willful attempt of taking influence of other players' moves, cannot be covered by rationality in a game-theoretic sense. Furthermore, game-theoretic research has to face the fact that – even in well-controlled and seemingly easily modeled situations – human players sometimes just don't behave like the solution concepts suggest. An example is the *centipede game* in Figure 2.7 (first introduced in [89]). Two players take turns and have the choice to either increase their utility by one and the game continues or to increase their utility by two and the game instantly ends. At some point the game ends anyway, because the resource that is depleted for

the utility is limited. In the penultimate possible move (the last one would

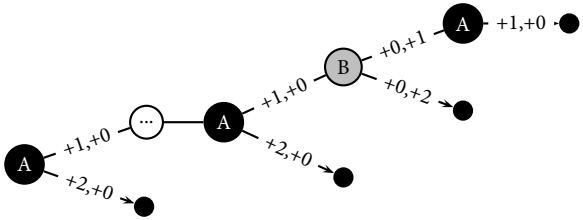


Figure 2.7: *Centipede game.*

completely exhaust the resource), the player better chooses a utility increase of two. The alternative to take one and leave the last unit of the resource to the other player would be irrational. In the preceding move, the other player can anticipate this move and – as he or she will not have another go anyway – better chooses a utility increase of two. Since this would again end the game, the other player better ends the game in the move before, and so on and so forth. This principle is called *backwards induction*: starting with an outcome, the move sequence that leads to this outcome is anticipated move by move. If the moving player has a higher utility when it deviates from the path that leads to the observed outcome, the outcome is excluded as a solution of the game. When repeated for all outcomes, this produces a set of so-called *subgame-perfect* Nash equilibria [92]. In a regular Nash equilibrium, no player has an incentive to change its strategy because the actions of the other players (according to their strategy) in the remaining part of the game would lead to an unattractive outcome. Occasionally, after one player changes its strategy, other players also have an incentive to deviate from their original strategy for the remaining subgame. The affected moves of the original strategy are called *incredible threats*, since they are not actually played when the player is confronted with the according decision. In a subgame-perfect Nash equilibrium, no player has an incentive to change its strategy at any point of the game. While this produces plausible results in many cases, the only subgame-perfect

equilibrium of the centipede game suggest that the first move ends the game. In reality, though, players keep going until the end of the game is in sight. At first glance, limited far-sightedness of the players could be an explanation. However, knowledge about the game mechanics by experience or explanation does not change the finding significantly. Apparently, both players keep the game alive in silent agreement as they see a high potential benefit in it. The risk of an uncooperative move by the other player seems to be low in the beginning. But it increases as the game proceeds and the utility from ending the game becomes significant in comparison to the remaining resources. Such behavior is covered in *repeated games*, where the players also consider future utilities of a later iteration of the game in their decision. Chapter 3 utilizes the concept of subgame-perfect equilibria to investigate instance pricing. A repeated game is used in Section 6.6.

Limits of Game-Theoretical Models

A proper modeling is necessary in order to obtain realistic game solutions. When possible moves, information status and utilities are misjudged, reliable conclusions for the modeled situation cannot be drawn. The game theoretic analysis implies that these aspects are *common knowledge*, which means that every player has the same understanding of the game. This also includes the knowledge that the game rules are common knowledge.

The design of game-theoretic models is not trivial. How much these models simplify real-world scenarios, which usually are of high complexity, should not be misjudged. This limits the use of game theory to rather small interactions. This holds also true because additional players add dimensions to the game's utility space. This makes the analysis more difficult. Nevertheless, game theory has become an important and well-established area especially in economics and sociology. The prisoners' dilemma is an example of why this theoretic approach is so successful: Game theory provides an explana-

tion of behavior that appears to be unreasonable from an outside perspective. Many interaction scenarios could not be described or comprehended before. Game theory even unravels the ostensible conflict of egoistic behavior – which might have evolved socially or genetically – and the widespread phenomenon of cooperation and supposed altruism: it can explain cooperation as a logical consequence from the dynamics of egoistic actors [12].

Note that non-cooperative game theory does not imply that there cannot be cooperation. The important part is that it is a result of a tacit understanding in which the cooperative act is everybody's best individual choice. This does not include an active form of cooperation where several players make an agreement on how they split up their combined utility. Such games are investigated in their own domain of cooperative game theory.

2.2.3 Evolutionary Game Theory

The achievements of game theory have led to an adoption of this approach in a wide range of disciplines. They usually follow the rationale that has been described so far. In biology, though, *evolutionary game theory* became a successful variety since Charles Lewontin first adopted the concepts of game theory for evolutionary dynamics [68].

Usually, game theory argues that players make rational choices based on comprehension of the game. Evolutionary game theory, on the other hand, tries to apply game theory on situations where the players lack the ability to make such deliberate choices. The players of an evolutionary game are not individuals, but there is one set of individuals that faces itself in a symmetric game. This set is called *population*. The population may feature a variety of behaviors (strategies), but the behavior of each individual is fixed. According to the composition of the population, it features a specific behavior distribution. This distribution is analog to a mixed strategy. It is not chosen, though, but

can change only over *generations* with the composition of the population. In biology, the term generation can be taken literally: evolutionary game theory is applied to animal populations, where behavior is regarded to be genetically determined. Some behavior may be more successful than some other and the fraction of individuals that feature this better behavior is hence increasing from one generation to another. Successful means anything that leads to increased reproduction (e.g. a larger amount of offspring). Instead of speaking of utility, evolutionary game theory uses the expression *fitness*. The fitness of a strategy is never evaluated regarding a potential future outcome of the game. Instead, it depends on the current population, which determines the competitive environment of each individual. A certain instruction called *replicator dynamics* defines the transformation of a population from one generation to another based on the current fitness values.

Over several generations, dominated strategies may vanish completely and other strategies may reach stable shares in the population. A stable strategy distribution is very similar to a Nash equilibrium. But Nash equilibria can sometimes be upset by coincidental strategy changes (e.g. caused by mutations or invading individuals). Evolutionary game theory hence defines a subset of Nash equilibria called the *Evolutionarily Stable Strategies (ESSs)* by asking for an additional stability condition: Any strategy m other than the equilibrium strategy n has to have a lower fitness than n in an environment following strategy m [96]. This refinement sorts out equilibrium solutions where some strategy distribution is able to supersede the equilibrium strategy distribution. The higher fitness of an ESS prohibits a growing share of other strategies in the population. As an example, Figure 2.8 shows the sex ratio of a mammal population based on Fisher's principle [35]. (The basic argument was already made in [26] and a first game-theoretic modeling of sex ratios can be found in [46].)

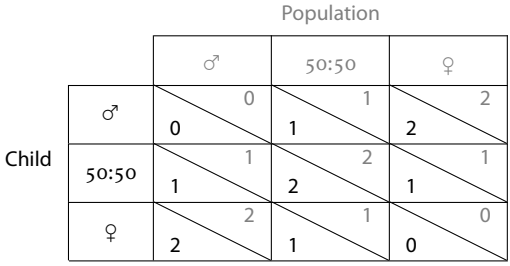


Figure 2.8: Another *battle of the sexes*: sex ratio in evolution.

Supposed there is a genetic predisposition to produce children of a certain sex. The predisposition can be to either have a higher probability of producing male children or a higher probability to produce female children or to produce either sex with equal probability. This is modeled in three strategies different strategies that represent the child's sex. When the child meets another individual in the population, there is a certain chance that this individual features the opposite sex and they can have children themselves. The fitness value represents the relative success of this reproduction. A couple that features a preference to produce a certain sex increases the share of this sex and the genetic predisposition to produce this sex in the following generation.

Assuming an imbalance in sex ratio, individuals of the majority sex have a lower chance to meet a partner of the minority sex and eventually have children themselves. In consequence, the predisposition to produce the majority sex has a lower fitness than the strategy to produce both sexes with the same probability. Accordingly, the strategy of an equal sex ratio is an ESS. It is restored after a disturbance that causes an imbalance and also cannot be invaded by any preference to produce a certain sex. Notice that the population strategy of a 1:1 ratio can either be constituted by a pure strategy where all individuals of the population produce male and female children with the same

probability or by a mixed strategy where the preferences to produce male and female children are present in equal shares.

In contrast to a simple Nash equilibrium, which can be regarded as stable because nobody has an incentive to leave it, stability is a dynamic property in evolutionary game theory. While the name of evolutionary game theory and its terms may suggest that it can only be used in a biological context, this is not true. It can also be applied in economic settings, for instance, where the population represents market shares [48]. Economic success supersedes reproduction success in this regard. Also, a generation is then better regarded as the discrete strategy distribution of a population at a specific point in time of a continuous development. Chapter 5 uses evolutionary game theory in order to investigate the competitiveness of physically separated processing and storage facilities.

"(...) replace up-front capital infrastructure expenses with low variable costs (...)"

– Amazon website

3 Pricing & Usage Dynamics

3.1 Introduction

When a client considers cloud computing, a variety of factors like privacy concerns and strategic decisions have to be reckoned with the particular case. In particular, costs are a key factor in the decision process. Case studies that ask whether or not processing in the cloud is feasible usually only regard current prices for cloud services. Since the cloud market develops, assuming constant prices is insufficient for long-term decisions. This chapter abstracts from current market prices and investigates the interaction of cloud provider and clients from an analytical perspective. A general understanding of how providers and clients potentially benefit financially from IaaS can help clients to appraise price uncertainty in strategic resource planning decisions. Providers gain insights on how pricing and charging models affect service usage.

The analysis focuses especially on the combined use of cloud services and an own data center, which offers a variety of possibilities how clients may split up their processing demand. While cloud service prices are most likely considered in the resource allocation decision of a client, it is unknown how this interrelation affects future cloud pricing.

Market dynamics depend on provider and client behavior. By contrasting the possible actions of these market actors, game theory can identify stable

market situations that suggest likely or advisable behavior. This chapter proposes a game-theoretic model and determines its equilibria in order to estimate future pricing and expected usage of IaaS in hybrid cloud scenarios. It further discusses the impact of factors like load distribution and economies of scale on the model. Also, the effects of a simultaneous offer of reserved instances is explored.

This chapter is organized as follows. The contribution of this chapter is put into relation to other research in Section 3.2. A market model for on-demand cloud infrastructure is developed throughout Section 3.3. In Section 3.4, the model is applied to an example case. Section 3.5 discusses the impact of a reserved charging option on the market. A general discussion of research results is presented in Section 3.6, which also concludes the chapter.

3.2 Related Work

Several publications deal with the suitability of cloud services as a substitute for on-site corporate IT. Guidelines for the decision process like [57, 65] usually include a financial comparison of feasible solutions. Calculation models for *Total Cost of Ownership (TCO)* of a data center [23, 58] can be taken as a basis for this. There are also ready-to-use calculators [6] for a direct comparison of expected costs based on specified demand. A cost model specifically for hybrid clouds is provided by [54]. Contrary to such case-based comparisons from a client's point of view, our approach allows to draw general conclusions on pricing and usage dynamics from a market model that considers the provider's perspective as well.

There also is game-theoretic research on cloud computing. Many publications focus on algorithmic solutions in resource management, e.g. [70, 105]. Several recent papers also discuss cloud instance pricing, though. For mar-

kets where providers compete in quality of service and price, the existence of a Nash equilibrium solution has been shown for a duopoly [31] and for n competitors [80]. In [9], an evolutionary approach is used to determine stable pricing. The authors in [51] study how the competition amongst clients for resources affects pricing. Instead of exploring instance price in a competitive environment, this chapter examines price and service utilization in hybrid clouds when the provider is in a monopoly position.

The negotiation of cloud instance pricing is approached by [107] in a 'one provider – one customer' bargaining problem. This chapter contributes a model where one provider serves many customers and the price is not bargained but fixed. Whether a provider should better provide instances on demand and at a fixed price or offer them in a spot market is discussed in [1]; a combined approach is also explored. [104] explores the distribution of resources on reserved, on-demand and spot market instances that maximizes provider revenue. We contribute similar research by investigating whether a provider should offer reserved instances, on-demand instances or both and what prices for the different instance types maximize provider profit.

3.3 A Game-Theoretic Market Model

3.3.1 Setup

A game-theoretic model of an on-demand infrastructure cloud market is suggested in the following; the goal is to estimate future pricing. The model is set up as one player being a provider that offers an on-demand computation instance. Such an instance provides capacity for processing and storage like a physical server and fees apply only when the instance is in use. The other player is a client that may utilizes the offered product. An *extensive form game* (Section 2.2) is used since the provider makes an offer and subsequently the

client is free to accept it or not. Figure 3.1 shows the proposed game model. The provider chooses a price and subsequently the client is free to use the service at these terms at any amount in combination with an own infrastructure in order to meet its processing demand.

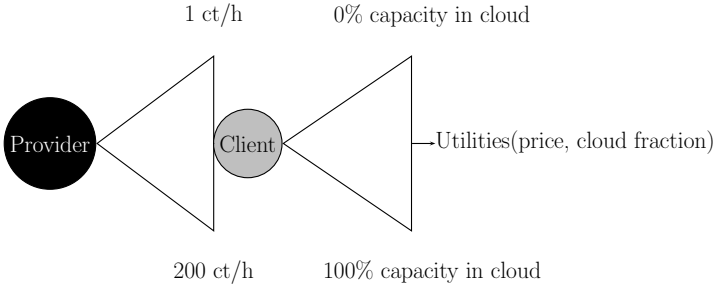


Figure 3.1: A market model for IaaS using continuous strategies.

We assume that the provider tries to maximize profit and the client tries to minimize expenses. Hence, utilities are based on client costs and provider profit. How the model can be solved based on these utilities is discussed in the following.

At a constant utilization of the cloud service, there is an incentive to raise the instance price as this increases the profit of the provider. This incentive is not given when a further price raise would result in lower profit. This can be the case, for example, when the client stops using the cloud. The client may follow the strategy to use the cloud when the price is lower than some $p_{\text{threshold}}$ and to build a data center when the price is higher than $p_{\text{threshold}}$. This client's strategy is in a Nash equilibrium (Section 2.2) with the provider's choice of a price that is just below $p_{\text{threshold}}$. A Nash equilibrium exists for any threshold price $p_{\text{threshold}}$ amongst many others.

The client will accede to the agreement, though, whenever a price is offered that actually causes lower costs in the cloud. As a consequence, any equilibrium where the client has an incentive to change its strategy during the game can be considered a non-credible threat. For example, suppose a specific cloud service price at which the cloud solution is a lot cheaper than an own data center to the client but the provider still makes positive profit. Further, the client follows the strategy to use the cloud service up to this price and to use an own data center if the cloud service is more expensive. Then this client's strategy is in equilibrium with the provider's strategy to offer exactly this price. Nevertheless, once the provider actually offers a higher price, the client might still benefit from the cloud service and prefer it over the data center option. Since the client will then change its strategy accordingly, the original equilibrium strategy is non-credible. As a consequence, unilateral change (changing the price) is safe for the provider when the client subsequently changes its strategy as well. Non-credible equilibria can thus be safely ignored. They can be sorted out by asking for *subgame perfection* which means that strategies in equilibrium have to remain best responses throughout all possible in-game situations (Section 2.2).

Clients are expected to choose different amounts of cloud instances based on the provider's offer. What fraction of a client's demand is met in the cloud depends on what combination of cloud and own infrastructure causes the lowest costs to the client at a given price. The price that offers the highest profit to the provider is in a subgame-perfect Nash equilibrium with the cloud fraction that belongs to the cheapest solution to the client at any given price.

Throughout the following sections, the utility functions of the game model are developed. Section 3.3.7 presents the final utilities and a calculation of the equilibrium solution.

3.3.2 A Simple Cost Model

Several calculation models for comparing the cost structure of cloud services and local data centers exist [6, 23, 58]. The cost models underlying these case studies can be easily generalized into two linear relationships.

Let us consider data center costs as the product of a number of servers n that is necessary to meet the maximum demand of a client and a constant c that includes all annual amortization and operation costs for one server (Equation 3.1).

$$\mathrm{DC}_{\mathrm{costs}}(c, n) = c \cdot n \tag{3.1}$$

Let us further consider cloud service costs as also depending on a number of servers n as a measure for the capacity that is used at a maximum. Unlike a data center, on-demand cloud instances can be scaled according to demand. We hence assume that the average number of instances in use correlates with the average workload a ($0 \leq a \leq 1$), which is a fraction of the capacity that is needed at a maximum. Factor e defines how many instances are equivalent to one server and p is the cloud instance price per hour. The annual costs of a cloud service are hence calculated as a product of these factors and the number of hours per year, as presented in Equation 3.2.[1]

$$\mathrm{Cloud}_{\mathrm{costs}}(p, e, n, a) = p \cdot e \cdot 24 \cdot 365 \cdot n \cdot a \tag{3.2}$$

3.3.3 Pricing Without Hybrid Clouds

When hybrid clouds are not possible, a client has to decide whether to utilize the cloud or to build an own data center for its *entire* demand. Considering

[1] More general cost models, e.g., a model where $\mathrm{Cloud}_{\mathrm{costs}}(a, n) = c_1 an + c_2 a + c_3 n$, are easily conceived. We restrict the discussion in this chapter to a linear model; yet our approach should carry over to such affine cost models as well.

the definitions of data center and cloud costs in Equations 3.1 and 3.2, we can expect the demand to stay constant as long as cloud pricing results in lower costs in the cloud compared to an own data center. For any average workload a, there is a break-even price $p_{\text{break}-\text{even}}$ at which the cloud and an own data center have equal costs. This can be expressed as a function $p_{\text{break}-\text{even}}(a)$ that is decreasing over workload a when data center costs and instance equivalent are considered constant (Equation 3.3).

$$\begin{aligned} \text{Cloud}_{\text{costs}}(p_{\text{break}-\text{even}}(a), e, n, a) &= \text{DC}_{\text{costs}}(c, n) \\ \Leftrightarrow p_{\text{break}-\text{even}}(a) &= \frac{c}{e \cdot 24 \cdot 365 \cdot a} \end{aligned} \tag{3.3}$$

A price higher than $p_{\text{break-even}}$ makes the cloud option financially unattractive for a client with the given workload. Thus, pricing unsurprisingly affects sales volume as well as profit margin. Long-term pricing will in all likelihood maximize the product of these factors, because providers want to maximize their profit. Without hybrid clouds, a provider can max out $p_{\text{break-even}}$ and clients are eventually left without any financial benefit from the cloud at all. The proposed game-theoretic model explores how this is affected when the cloud and own infrastructure can be combined and complex load situations apply. How to determine the best price from a provider's perspective in such a scenario is discussed in the following sections.

3.3.4 Combining Cloud and Data Center

Projects like *Eucalyptus* [33] and *Openstack* [79] provide implementations for a local deployment of cloud services. A client might want to use such software to combine a *private cloud* (build an own data center providing cloud services) and the *public cloud* (Internet cloud services) to meet its demand. Such combinations are called *hybrid clouds*.

Usually, some capacity is always in use (*base load*) and some capacity is not used continually but only part-time (*peak load*). The necessary total processing capacity at a maximum is called *maximum demand* in this chapter. The demand that is produced by base load over time is referred to as *base volume*, all exceeding demand is called *peak volume*. Accordingly, base volume and peak volume sum up to the total processing demand of a client over time. Peak volume is more expensive to self-provide than base volume, as costs are amortized only over the time the necessary capacity is actually used. Thus, building a smaller data center to meet base load and buy instances from the cloud to meet peak load could be a sensible choice. This also means that higher prices can be tolerated for public cloud instances.

From a client's perspective, all options seem to be reasonable in their own range of cloud pricing. Instead of a break-even price where meeting the entire demand in the cloud is as expensive as an own data center ($p_{\text{break-even}}$, Section 3.3.3), there are two other break-even prices: First, a $p_{\text{low}} < p_{\text{break-even}}$ where the pure cloud solution costs the same as a hybrid cloud in which an own infrastructure is catering only to base load. And second, a $p_{\text{high}} > p_{\text{break-even}}$ where the hybrid cloud is cheaper than an own data center covering all demand. Complete outsourcing is most attractive when the service is cheaper than p_{low}, the hybrid cloud setup is the best option between p_{low} and p_{high}, and not using the cloud service at all is the cheapest option when the instance price exceeds p_{high}. The more peak volume the client demands (at constant maximum demand), the less do p_{low} and p_{high} differ. Figure 3.2 shows an example where the different break-even prices are pointed out. The client has an incentive to change the solution when p_{low} or p_{high} is exceeded. These solution changes come along with a drop in demand and thus in the associated profit to the provider. Hence, there are two pricing candidates for profit maximization, when hybrid clouds are regarded: the lower price is right below p_{low}, the higher price is right below p_{high}. The costs of cloud and hybrid

cloud depend on the average workload a. Accordingly, p_{high} depends on the load profile of the client, which is further discussed in Section 3.3.5. However, this is not the case for p_{low}, which depends only on data center costs of the client and the relative capability of an instance (Equation 3.4). (This means that although the straights for cloud and hybrid cloud change with the load profile, they still intersect at the same price p_{low}.)

$$\begin{aligned} \text{DC}_{\text{costs}}(c, n) &= \text{Cloud}_{\text{costs}}(p_{\text{low}}, e, n, 1) \\ \Leftrightarrow p_{\text{low}} &= \frac{c}{24 \cdot 365 \cdot e} \end{aligned} \tag{3.4}$$

Revenue from covering all volume at p_{low} might be higher or lower than revenue from covering peak volume at p_{high}. The load profile determines break-even costs of the cloud and the hybrid cloud solution (Figure 3.2) and hence also determines which price is more appealing to offer.

3.3.5 Different Load Profiles

It is insufficient to limit hybrid clouds to the single combination where an own infrastructure is used for base load and the public cloud is used for all peak volume. Load profiles can be described by a *Complementary Cumulative Distribution Function (CCDF)*. Function time(x), where x is some amount of load, returns the amount of time where the data center is under a load of x or more (Figure 3.3). Base load (here: 30%) is always in use, higher utilization occurs more and more rarely. The example load profiles feature 50% average utilization of either all capacity (*convex*) or of all capacity exceeding base load (*even*). A load curve can feature a smooth transition between peak and base load. Thus, several capacities of a data center might be reasonable for a hybrid cloud. Higher instance prices justify a larger data center when extra capacity is

used cost-effective. This means that it would be more expensive to substitute the extra capacity with public cloud instances.

All hybrid cloud possibilities are covered by the second continuum of the game model in Section 3.3.1. It allows the client to choose any fraction of overall capacity that is met by a data center (private cloud). This fraction is called DC_{frac} in the following. It ranges from an exclusive use of an own data center to the utilization of cloud instances for all demand; a value of 0.4, for instance, means that an own data center is capable of meeting 40% of maximum demand while public cloud instances are used to meet any further demand. Cloud capacity for the load that exceeds DC_{frac} is never used for a higher frac-

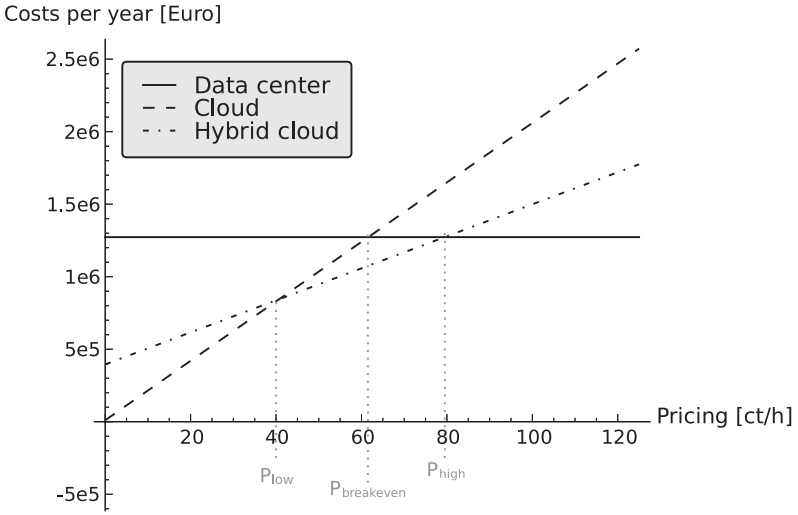


Figure 3.2: An example for costs of data center, cloud and hybrid cloud over on-demand instance price. From a client's perspective, all options can be reasonable.

tion of time than time($\mathrm{DC}_{\mathrm{frac}}$). Accordingly, the cost-optimal $\mathrm{DC}_{\mathrm{frac}}$ for any on-demand price is given when data center capacity costs the same as cloud capacity that is only used for time($\mathrm{DC}_{\mathrm{frac}}$). For example, when cloud capacity is double the price as data center capacity, it is best used for capacity that is only required half the time or less, i.e. time($\mathrm{DC}_{\mathrm{frac}}$) $= 0.5$. The according $\mathrm{DC}_{\mathrm{frac}}$ can be calculated as presented in Equation 3.5. We define the inverse function time$^{-1}(z)$ as zero for any input value $z \geq 1$ (occurs at instance prices $p \leq p_{\mathrm{low}}$ where cloud is always cheaper than data center).

$$\begin{aligned} &\mathrm{DC}_{\mathrm{costs}}(c,n) = \mathrm{Cloud}_{\mathrm{costs}}\left(p,e,n,\mathrm{time}(\mathrm{DC}_{\mathrm{frac}}(\mathrm{p}))\right) \\ \Leftrightarrow\ &\mathrm{DC}_{\mathrm{frac}}(p) = \mathrm{time}^{-1}\left(\frac{\mathrm{DC}_{\mathrm{costs}}(c,n)}{\mathrm{Cloud}_{\mathrm{costs}}(p,e,n,1)}\right) \end{aligned} \tag{3.5}$$

Since the client reduces the cloud share of its demand when the price is increased, there is not just one break-even price p_{high} of a single specific hybrid

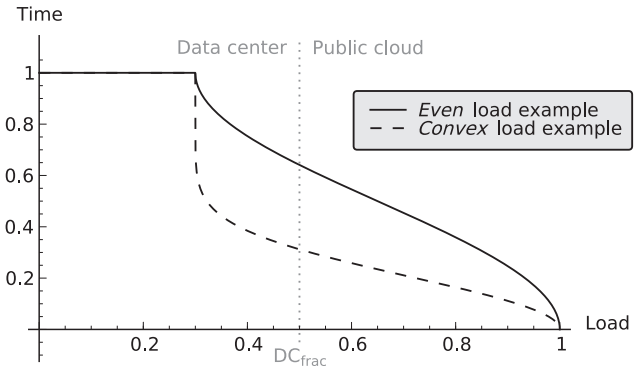


Figure 3.3: Two example load profiles.

cloud solution that is a credible price candidate. A different price causes a different combination of the public cloud and own infrastructure to be the cheapest for the client. The best price hence has to be determined by backwards induction, as presented in Section 3.3.7.

3.3.6 Provider Profit

So far, only the revenue of the provider has been considered. Although initial growth might be a business objective, revenue as such is to no avail if the business is not profitable in the end. Thus, it is important to include provider costs into the model to have its utilities based on profit. The provider actually has to build a data center itself to provide the service. In order to provide a service of the same capabilities that a client's local data center would feature, the provider has to operate equivalent hardware. In consequence, data center costs of the provider can be estimated like data center costs of a client in general. Without some cost advantage in operation, a profitable offer at p_{low} would be impossible.

For a huge provider, there is a cost advantage due to economies of scale in the first place. Advantages of location like cheaper power or building costs might provide further benefits to the provider. On the other hand, renting out instances on demand comes along with unpaid idle time. To compensate for the investment, time under utilization would have to be more expensive than p_{low}.

Instances can be overbooked by statistical multiplexing, though. There is no need to operate the same amount of servers as a client would have to, but only the amount to meet average demand (in the best case). By doing so, the provider runs the risk of not fulfilling the service level agreements; the probability depends on how much peaks are correlated in time. Several client's peaks may occur at the same time. Suppose the chance that a client actually

demands an instance has a uniform distribution over time, then the larger the number of clients, the less variation in overall average demand can be expected (law of large numbers). Demand correlation can be an issue, though (Section 3.6.5).

The costs of the provider's data center are called $\mathrm{DC}_{\mathrm{costsPr}}$. Following what is said above, they are calculated as presented in Equation 3.6. Similar to what a client would pay for its own infrastructure (Section 3.3.2), costs are based on a number of servers n and data center costs per server c. The required capacity is reduced to what is needed for average load a. There has to be a reserve for the variation in total demand the provider meets, though, which is defined by a coefficient v. EoS defines the provider's costs savings due to economies of scale. The factor is defined as the fraction of data center costs of the provider in comparison to the costs of the client for a data center of the same capacity.

$$\mathrm{DC}_{\mathrm{costsPr}}(c, n, a, \mathrm{EoS}, v) = \mathrm{DC}_{\mathrm{costs}}(c, n \cdot a \cdot v) \cdot \mathrm{EoS} \tag{3.6}$$

As explained earlier, v goes to 1 when a provider serves a very large number of clients of huge diversity. Compared to a smaller provider with fewer clients, a huge provider accordingly requires less reserve capacity to cope with demand variation. The smaller provider has to pass on the cost of extra capacity to actually sold instances. As a consequence, the huge provider can offer it's service at a relatively low price (e.g. p_{low}) and make a profit, while the same price might not yield a profit for the smaller provider (p_{low} becomes unprofitable when $v > (1/\mathrm{EoS})$).

3.3.7 Utilities & Subgame Perfect Nash Equilibrium

Based on the preliminary work of the preceding sections, the final utility functions of the model are given in Equations 3.7 and 3.8. The utilities are based

on revenue and expenses that apply at the chosen price and service usage: Provider utility is it's profit, client utility is the negative overall cost of meeting processing demand. Again, n is the number of servers necessary to meet maximum demand, e is the number of instances necessary to substitute one server and c is data center costs per server. The move by the provider is to choose an instance price p. The move by the client is to choose $\mathrm{DC}_{\mathrm{frac}}$, the fraction of maximum demand that is handled on own infrastructure. The average workload a of demand that is met in the cloud changes with $\mathrm{DC}_{\mathrm{frac}}$ since only the exceeding demand adds to the expected period of time that a cloud instance is in use. The average workload is accordingly determined by an integral over the load distribution on the interval $[\mathrm{DC}_{\mathrm{frac}}, 1]$. All other assessments are presented in the preceding sections: $\mathrm{DC}_{\mathrm{costs}}$ and $\mathrm{Cloud}_{\mathrm{costs}}$ are defined in Section 3.3.2, $\mathrm{time}(z)$ is a load distribution as described in Section 3.3.5 and $\mathrm{DC}_{\mathrm{costsPr}}$ can be found in Section 3.3.6.

$$\begin{aligned}\mathrm{Client}_{\mathrm{utility}}(p, c, e, n, \mathrm{DC}_{\mathrm{frac}}) = &-\mathrm{DC}_{\mathrm{costs}}(c, n \cdot \mathrm{DC}_{\mathrm{frac}}) \\ &- \mathrm{Cloud}_{\mathrm{costs}}\left(p, e, n, \int_{\mathrm{DC}_{\mathrm{frac}}}^{1} \mathrm{time}(z)\mathrm{d}z\right)\end{aligned} \tag{3.7}$$

$$\begin{aligned}\mathrm{Provider}_{\mathrm{utility}}(p, c, e, n, \mathrm{DC}_{\mathrm{frac}}) = &\ \mathrm{Cloud}_{\mathrm{costs}}\left(p, e, n, \int_{\mathrm{DC}_{\mathrm{frac}}}^{1} \mathrm{time}(z)\mathrm{d}z\right) \\ &- \mathrm{DC}_{\mathrm{costsPr}}\left(c, n, \int_{\mathrm{DC}_{\mathrm{frac}}}^{1} \mathrm{time}(z)\mathrm{d}z, \mathrm{EoS}, v\right)\end{aligned} \tag{3.8}$$

The equilibrium solution of the game can be determined by *backwards induction*: the profit at any possible price is determined respecting the expected service usage by the client; the price that coincides with maximum profit is the

price in equilibrium. According to Equation 3.8, the provider utility depends on the amount of usage that can be expected under this pricing (DC_{frac}). For each price, a different DC_{frac} applies and can be determined with Equation 3.5. We call the price where the provider's utility is at its maximum p_{best} (Condition 3.9).

$$\begin{aligned} &\partial_p \text{Provider}_{\text{utility}}(p_{\text{best}}, c, e, n, DC_{frac}) = 0 \quad \text{and} \\ &\partial_p^2 \text{Provider}_{\text{utility}}(p_{\text{best}}, c, e, n, DC_{frac}) < 0 \end{aligned} \tag{3.9}$$

p_{best} might be equal to p_{low}. The game features a subgame-perfect Nash equilibrium within the combination of instance price p_{best} and the client combining public and private cloud services in a cost-optimal split-up $DC_{frac}(p_{best})$. p_{best} equals p_{low} at a minimum.

3.3.8 Client Demand Aggregation

The former considerations are based on the demand of one client. To make a statement on probable long-term pricing, the dynamics between provider and all potential clients have to be considered. Since the profit maximizing price of different clients usually differs, it is not possible to maximize profit from all individual clients at the same time (assuming that the provider offers the same price to all clients in the market). Instead, the provider has to consider the expected usage of the entire market (best responses of all individual clients) in order to determine the profit maximizing price. To achieve this, the single client of the presented model can be replaced by an aggregation of all potential clients: a meta-client with a distinct demand. The maximum demand of the meta-client is the sum of all individual maximum demand.Its load can be described by the antiderivative of the convolution of all individual load functions' derivatives. The result is a CDF for the entire market that needs to be

subtracted from 1 to get a CCDF as it is used in Section 3.3.5. Now, the equilibrium price does not maximize the profit that the provider yields from any individual client, but p_{best} based on this new load curve is the best trade-off that is possible with a single price for all clients. Research on the distribution of load profiles and associated demand volume is needed to tell how the overall demand of the market would look like. However, with the presented aggregation of demand, what was said about the single-client-situation can be generalized to the whole market.

3.4 Future Infrastructure Cloud Pricing

3.4.1 Case Study

The German IT magazine *iX* published a case study for a hypothetical company in which the TCO of a new data center was compared to the costs of a co-location setup and the use of Amazon's *EC2* service [23]. In the study, costs per year for an owned data center consist of investment cost amortization and running costs. Investments are acquisition costs for server and network hardware and operation system licenses (3 years write-off) as well as infrastructure and building costs (15 years write-off). Running costs are maintenance, power, administration, and data transfer. This results in annual data center costs of approximately € 7150 per server. The processing capability of one of the considered servers is regarded as equivalent to two EC2 instances and the study uses an instance price of € 0.22/h.

The underlying cost model and the client case of this study are of course debatable. Nevertheless, we use it as an example application of our model and configure $c = 7150$, $e = 2$ and $n = 180$ for the numerical calculations in this section. The instance price is not derived from the case study, of course, as it is chosen by the provider in our model. We also use load distri-

butions according to Section 3.3.5. Further, a demand variation coefficient of $v = 1$ is assumed, keeping in mind that this only represents the best case for the provider. Despite the use of EC2 in this example, $\text{Cloud}_{\text{costs}}$ should be considered as a representative for all kinds of cloud offerings, not necessarily restricted to the Amazon cloud as such.

To keep things simple, co-location (buy servers and lease facilities and administration for operation) is not included as an option like it is in [23]. It need not be regarded as essentially different from an owned data center (buy servers and build facilities, employ administration personnel for operation) and is briefly discussed in Section 3.6.

3.4.2 Provider Profit Estimation

The profit margin of the provider depends not only on the load distribution, but also on the economies of scale of the provider (Section 3.3.6). In the following, we try to estimate what scale economies are realistic based on the case study's cost model.

In [23], the following factors are considered. Network hardware is assessed at 20% of server cost, annual maintenance at 10% of server and network hardware costs. Power consumption of a server is estimated at 50% of its specification, consumption of network gear at 44% of server consumption. For the total data center consumption these values are multiplied by a PUE (Power Usage Effectiveness) factor of 2.5 and assessed at €0.1/kWh. The building and infrastructure investments are measured at € 2024/m^2 (2.84 m^2 per rack) and € 16200 per kW hardware power consumption (referring to uptime institute [100]). Administration costs are set to € 73000 per administrator and year. Each administrator is capable of covering 80 servers. Data transfer is included at a flat rate of € 400/month.

According to James Hamilton from *Amazon Web Services*, the PUE of a large data center (50000+ servers) is between 1.2 and 1.5 and in comparison to a mid-sized data center (∼1000 servers) admin costs can be reduced by factor 7 due to automatization [45]. Compared to the case study's calculation, this reduces average costs per server by about 18%. A little over 20% would be theoretically possible with full automatization (no admin costs) and an ideal PUE of 1. This corresponds to $\text{EoS} = 0.8$ in our model.

3.4.3 Applying the Model to the Case Study

A client's best response on certain pricing remains unaltered by different load profiles up to instance costs that – when continually paid – correspond to the costs of a data center equivalent: p_{low}. A higher price causes a drop in demand because base load (lots of volume) is handled by owned data centers. Outsourcing processing peaks generates decreasing demand over price. Figure 3.4 shows an example calculation where the model is applied to the case study from Section 3.4.1. It shows the client's cost for the best hybrid cloud solution as well as provider revenue (public cloud share of hybrid cloud costs) and profit (revenue minus production costs) over instance price. At $p_{\text{low}} \approx$ €0.40/h, the first peak appears, the later local maximum of provider profit happens at $p_{\text{best}} \approx$ €0.60/h.

How exactly the demand wears off over price depends on the load profile. As a consequence, load defines whether revenues at higher prices retain a volume which allows exceeding profit compared to p_{low}. Profit at p_{low} is gathered from mass usage at a comparatively low margin. As discussed in Section 3.3.6, the data center cost advantage is also determined by economies of scale and expected demand variation. The higher the pricing, the less important do these factors become for the margin and thus a smaller economies of scale factor makes a pricing at p_{low} more likely (Figure 3.5).

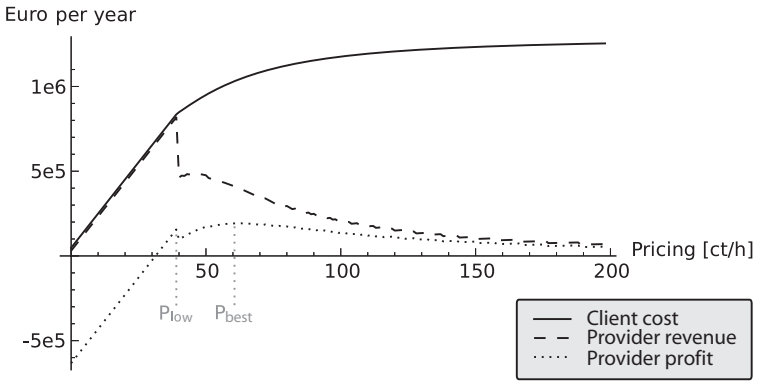


Figure 3.4: Example calculation of client cost for the cheapest hybrid cloud solutions over cloud instance price and associated provider revenue and profit (EoS = 0,8; $v = 1$).

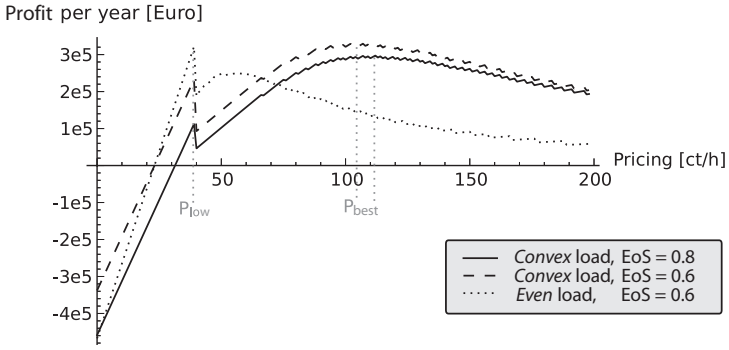


Figure 3.5: Expected provider profit at different economies of scale and load ($v = 1$).

Provided that the calculation assumptions of the case study are correct (especially that two EC2 instances are equivalent to one of the accounted servers), the EC2 service is not cost-effective at the instance price of €0.22/h that is used in [23]. Instead, a price raise to €0.40/h is likely in the future since client-owned infrastructure is generally more expensive than the cloud up to this price and the provider can hence increase the price without losing any demand. Assuming an average load distribution in the market that equals the *convex* load example (Figure 3.3), an even higher on-demand instance price of over €1/h is more profitable (Figure 3.5). It is therefore likely to be asked once the provider should achieve a monopoly position. At this instance price, a client that features *even* load would be best off building a hybrid cloud with a private cloud share of 75 % (Equation 3.5).

3.5 Including Reserved Instances in the Model

3.5.1 On-Demand and Reserved Instances

The presented model considers pricing and usage of *on-demand instances*, which means that they only have to be paid for the time in use. They can be a cost-saving substitute for a client's data center because the number of instances in use can be adapted to current demand. Due to statistical multiplexing, the hardware necessary to meet demand of all clients is fewer in a cloud environment than when the clients run their own data centers. This is the main benefit of the cloud next to economies of scale. These benefits are shared between client and provider. Section 3.3.4 shows that a provider may offers a high instance price in order to meet peak load only and clients still want to operate their own (but less capable) data centers in that case. Otherwise, when the provider prefers to meet all demand, the possibility of a hybrid cloud is a threat to the provider that ensures relatively low prices to the clients.

This chapter explores how this is affected when instances are not rented out on demand but where a different charging model is used: *reserved instances*.

On-demand and reserved instances feature the same technical specification at a different charging model. Unlike on-demand instances, reserved instances have to be paid for much longer time periods irrespective of actual use. On the other hand, the provider is liable for the availability of the instance during the reservation period. While we consider this liability difference between instance types from the provider's perspective in the following, the availability risk of on-demand instances to the clients is discussed later in 3.6.4. In order to give an incentive to make use of reserved instances, they have to be competitive with client-owned data centers. It is the cheapest option for the client to completely substitute an owned data center when instances are priced cheaper than their data center equivalent (p_{low}, Section 3.3.4). Offering both instance types at p_{low} does not make any sense because the on-demand option is more attractive to the client and less profitable for the provider. When on-demand instances are priced higher, though, it might be reasonable to offer both options. Based on the results from Section 3.3, the straight-forward solution appears to be to use p_{best} to target peak load and p_{low} for reserved instances: While this maximizes the profit from on-demand instances as before, reserved instances are preferred over data center capacity for all remaining demand, which yields additional profit to the provider.

3.5.2 Pricing On-Demand and Reserved Instances

Without reserved instances, clients keep more and more demand local as the on-demand price rises. All associated profit is lost for the provider. When the clients switch to reserved instances instead, the profit of these instances can compensate for the losses on the on-demand side. Hence, there might be a better on-demand instance price that exceeds p_{best}.

The game from Section 3.3 is modified in a way that the provider can now choose a combination of two prices, p for on-demand instances and p_{res} for reserved instances. The client now chooses a combination of both instance types. For the client, both instance types are interchangeable (availability provided) as they are technically equivalent. The higher the on-demand instance price, the higher the amount of reserved instances the client uses. Between both types exists a threshold T similar to DC_{frac}, which divides demand that is better met by on-demand or reserved instances (Equation 3.10).

$$\begin{aligned} \text{Cloud}_{\text{costs}}(p_{\text{res}}, e, n, 1) &= \text{Cloud}_{\text{costs}}\left(p, e, n, \text{time}(\text{T}(\text{p}_{\text{res}}, \text{p}))\right) \\ \Leftrightarrow \text{T}(p_{\text{res}}, p) &= \text{time}^{-1}\left(\frac{p_{\text{res}}}{p}\right) \end{aligned} \tag{3.10}$$

A downside of a high amount of reserved instances is a growing chance of an overbooking conflict (in case that the provider does not actually reserve instances for a single paying client, i.e. $a = 1$ for $\text{DC}_{\text{costsPr}}$). Capacities are rented out to several clients in the expectation that they are not used simultaneously. When the number of sold reserved instances ($T \cdot e \cdot n$) exceeds the available hardware, there is a certain chance that there is more demand at the same time than the provider can handle. The higher the amount of reserved instances, the higher is the chance that this happens. Exceeding on-demand instance demand can just be ignored. With reserved instances, on the other hand, the provider is unable to provide an already paid service and has to pay a contract penalty fine. Factor f is the annual fine that applies for one reserved instance that is paid but not available. The total amount of penalties that the provider has to pay is the product of this fine, the number of instances and the probability that a reserved instance exceeds available hardware. As presented in Section 3.3.6, this probability is zero in the best case (law of large numbers). Let us assume the worst case of complete demand correlation in the following.

Client and provider utilities from Equations 3.7 and 3.8 can be easily adapted to include both instance types and the penalty, as presented in Equations 3.11 and 3.12. Instead of data center costs, the client has costs for reserved instances, which on the other hand provide additional revenue to the provider.[2]

$$\begin{aligned}\text{Client}_{\text{utility}}2(p_{\text{res}}, p, e, n, T) = &-\text{Cloud}_{\text{costs}}(p_{\text{res}}, e, n, T)\\ &- \text{Cloud}_{\text{costs}}\left(p, e, n, \int_T^1 \text{time}(z)\text{d}z\right)\\ &+ f \cdot n \cdot e \cdot \int_{\int_0^1 \text{time}(z)\text{d}z \cdot v}^{T} \text{time}(z)\text{d}z\end{aligned} \tag{3.11}$$

$$\begin{aligned}\text{Provider}_{\text{utility}}2(p_{\text{res}}, p, c, e, n, T) = &\ \text{Cloud}_{\text{costs}}(p_{\text{res}}, e, n, T)\\ &+ \text{Cloud}_{\text{costs}}\left(p, e, n, \int_T^1 \text{time}(z)\text{d}z\right)\\ &- \text{DC}_{\text{costsPr}}\left(c, n, \int_0^1 \text{time}(z)\text{d}z, \text{EoS}, v\right)\\ &- f \cdot n \cdot e \cdot \int_{\int_0^1 \text{time}(z)\text{d}z \cdot v}^{T} \text{time}(z)\text{d}z\end{aligned} \tag{3.12}$$

The on-demand option is a good choice for peak load from a client's perspective. For the provider, though, offering this charging option next to reserved instances lowers its revenue. In Section 3.3, a higher on-demand instance price was providing a higher revenue but caused lower demand and

[2] Note that the model does not consider a client's damage from unmet demand (Section 3.6.4).

a trade-off between usage and price had to be determined. With reserved instances as the preferable fall-back solution for the client ($p_{\text{res}} \leq p_{\text{low}}$), all demand is met by the public cloud anyway. Whenever the client chooses on-demand over reserved instances because they are cheaper, the provider has to meet the same demand for less revenue. Nevertheless, when demand is correlated and there is some variance in demand, offering some capacity share on demand can reduce overbooking conflicts and contract penalty fines.

3.5.3 Reserved Instances in the Case Study

Figure 3.6 shows the provider's revenue and profit over on-demand instance price in presence of a reserved instance option that is offered at price p_{low}. Provider profit is presented both with and without considering contract penalties. Additionally, the profit in a market without reserved instances as presented in Section 3.4 is given for comparison. In the on-demand case there is a drop in demand when the price exceeds p_{low}. This is not observable when reserved instances are offered as well: instead of a private cloud, reserved instances are used for the base load. The more expensive on-demand instances become, the more reserved instances are used by the client. When the share of on-demand instances becomes too little, rising contract penalties may outgrow the additional revenue and limit the reasonable on-demand pricing to about €1.20/h in this example. Whether this is the case depends on how much demand is correlated, though. The curves for provider profit with and without penalty give the worst and best case of what might actually happen. Nonetheless, compared to an on-demand market, the best on-demand price p_{best} can be a lot higher with an additional reserved instance option since this does not drive clients out of the cloud. This may cause significantly higher overall costs to the client.

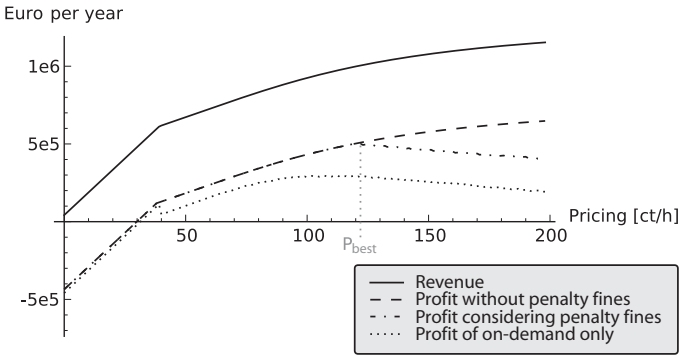


Figure 3.6: Provider revenue and profit in presence of a reserved instance option. Overbooking conflicts of reserved instances can cause penalty fines and reduce provider profit, which limits a reasonable on-demand price. ($p_{\text{res}} = p_{\text{low}}; v = 1; f = €1/h$)

3.5.4 Two-Part Tariffs

Next to on-demand and reserved instances, intermediary options with a combined charging model are also possible: *two-part tariffs*. Amazon currently uses this charging model for their *Light* and *Medium Utilization Reserved Instances*, for example (while the *Heavy* variant is reserved-only) [7]. The client pays a reduced charge for a long time period (similar to reserved instances) plus an additional fee for the time the instance is in use (similar to on-demand instances). This creates a whole spectrum of charging options. Starting at a reserved-only charge, the reserved price component shrinks while the on-demand price component rises until it finally ends up in an on-demand charge without a reservation fee. For n options, there are $n - 1$ thresholds that split demand in capacities of shrinking utilization which are each best met by a specific charging option.

Clients have better knowledge of their load profile than the provider. The distribution of instances over available charging options that they choose hence resembles an estimate of the actual utilization of the instances. This helps the provider to assess the necessary data center capabilities and prevents conflicts in overbooked capacities. The provider thus has an incentive to offer such a variety of charging options. Nevertheless, the additional options reduce provider profit in favor of the client (Section 3.5.2). Due to the better knowledge about the expected utilization, though, an on-demand charge becomes less important to maintain a sufficient level of non-reserved capacity in order to prevent conflicting usage and can accordingly be priced higher. Reserved instances with additional on-demand charges can encourage co-usage of several instance types and a client-owned data center when their pricing makes them the cheapest option for demand with a certain load. Two-part tariffs are further investigated in Chapter 4.

3.6 Discussion & Conclusions

This section generally discusses the presented model and briefly addresses a few questions that were raised earlier in the chapter. It also presents starting points for further research.

3.6.1 Effects of Hybrid Clouds

Cloud services might be cheap today, but things could look different as soon as cloud services are established and have to prove themselves as sustainably profitable. Their likely future pricing can be estimated based on the presented model and solid knowledge of the clients' load distribution in the market (load of the meta-client in Section 3.3.8). This prospect of cloud pricing should be

considered in today's resource allocation decisions. Each client can determine its individual, cost-optimal solution based on this price (Equation 3.5).

Not to use the public cloud service at all is never financially favorable to a client according to the market model (although the suggested solution might contain a very small on-demand public cloud share). In general, all clients have an advantage from hybrid clouds, but those with higher-than-average load gain less than the average client. The possibility of hybrid clouds only provides a significant improvement to all clients if it does not have to be implemented and instead all demand is met in the cloud: The provider has to reduce the price from $p_{\text{break-even}}$ to p_{low} in order to counteract the threat that the client operates a private cloud for base load. If the provider chooses to meet only peak load, though, it depends on the client's load distribution whether there are huge savings in comparison to an own data center.

The possible revenue for the provider is lower with hybrid clouds than without, where revenue can be as high as the costs of a client's data center. In a hybrid cloud scenario, the client operates own infrastructure for base load when the cloud price is p_{high}. Expenses for the cloud service are hence capped to the amount that necessary additional capacity would cost in order to meet all demand. At a cloud price p_{low}, all demand is met in the cloud. But as p_{low} is smaller than $p_{\text{break-even}}$, this means less revenue for the provider than without the possibility of hybrid clouds.

3.6.2 Effects of Reserved Instances

The existence of reserved instances in the market is primarily of advantage to the provider. Demand for reserved instances is very robust up to p_{low}, which is determined by the data center costs per server of the clients' hypothetical data centers. While a provider may offers instances at p_{low} in a on-demand market, reserved instances cause p_{low} to be a bad choice for on-demand in-

stances. In a monopoly, a provider would hence choose to offer reserved instances at a price p_{low} and on-demand instances at a lot higher price. This on-demand price is usually higher than p_{best} in a purely on demand-market. With reserved instances, there is no credible threat that the client moves capacity away from the public cloud. Hence, on-demand pricing is only limited by contract penalties for overbooking conflicts that potentially occur because of demand correlation (Section 3.5.2).

On-demand pricing that targets peak load instead of mass usage depends on good market knowledge and is thus easily misapplied. With reserved instances as an alternative, a higher on-demand price means higher revenue from reserved instances and only becomes a problem when overbooking conflicts occur. Two-part tariffs as discussed in Section 3.5.4 help to gain knowledge about the amount of utilization that can can be expected. From a financial viewpoint, the provider has an interest to discourage on-demand usage when peak load can be estimated well.

3.6.3 Effects of Economies of Scale and Market Form

Because p_{low} depends on good economies of scale to a large extent, it is important to understand which aspects are persistently cheaper on a larger scale. It is quite possible that technologies which provide a better PUE become available to smaller centers in the future as well, for instance. If production costs of instances become insignificantly cheaper than p_{low}, heavy utilized capacity is likely to be left to private clouds. The offer of such cheap reserved-only instances would be unattractive at such a low profit margin. Charging options that combine an on-demand and a reserved price component, on the other hand, can still be profitable.

An oligopoly appears to be the most preferable market form from a client's perspective. Undercutting competitors' offers keeps margins low, while pro-

duction costs are low due to the size of the providers. The whole aspect of competition needs a more in-depth analysis, of course, which is conducted in Chapter 4.

3.6.4 Availability Risk & Intangible Aspects

Section 3.5 pointed out that the availability of on-demand instances is not warranted. Since the capacity of the provider is limited, there is the risk that not enough on-demand instances necessary to meet a client's demand are available. This is not considered in the model. However, it could be included as an additional factor analog to the contract penalty fine to the provider: the client utility is lowered in case that an instance is not available. Depending on how crucial it is to a client that its demand is always met, it might prefer reserved instances over on-demand instances even if they are more expensive because the fine reimburses for the damage in that case or the change of an available instance is higher. In consideration of the higher risk of on-demand instances compared to reserved instances, it might occur counter-intuitive that on-demand instances should be more expensive (Section 3.5.2). Note that they can be a lot cheaper to a client despite their higher price when they are only used for short time periods.

With all debates on cloud-related security and privacy, risk in general obviously has a huge impact on decisions. The model does not account for any non-financial aspects. These aspects might be incorporated into the model in form of additional factors that lower (client) utilities similar to the availability risk. Rational quantification of risk is a difficult task, though, and very dependent on the individual case.

3.6.5 Remarks on Ultimatum Game, Demand Correlation and Colocation

It is important to easily combine instances from several sources (public and private clouds) in order that both parties benefit from the cloud. For some applications, data exchange between these sources might be an issue, though. Also, standards in use might not support hybrid clouds, e.g. by ensuring interoperability and migration. If a combination with arbitrary shares is not granted by the service, the best pricing is right below $p_{\text{break-even}}$ (Section 3.3.3). Although an exclusive use of the cloud minimizes overall idle time, which is good for environmental reasons, the provider has all the benefit. It is noteworthy, though, that the game setup (Section 3.3.1) is quite similar to the *ultimatum game* [41]. Results in experimental economics differ from theory regarding these games: a price that does not provide enough benefit to the client might be perceived as somehow inappropriate and is thus rejected.

Section 3.3.6 mentions that a smaller provider is more likely to have idle times than a huge provider due to suboptimal overbooking. Another reason might be that a huge share of clients might demand resources at the same time (e.g. at daytime in a single timezone). When the provider has to operate reserve capacity, this diminishes the profit margin. If savings due to economies of scale are exceeded by the additional costs, p_{low} becomes unprofitable. This not only means that a very large provider with clients scattered all over the globe can expect more benefits from the cloud than a small one serving regional clients. It also makes it very hard to establish a competitor in the market as massive investments can be expected.

In the case study that is used in Section 3.4, co-location is discussed as an option. It was omitted as an option to the client in the presented market model because it is quite similar to the data center option. A facility that houses several client's servers might be very large and in consequence there may be

economies of scale in favor of a co-location provider. Overbooking is not possible, though, as the servers are reserved for a specific client. Hence, the benefit that is shared amongst provider and clients is much smaller in comparison to cloud utilization. Providing inferior gain to both parties, the whole business model of co-location is challenged by the existence of cloud computing.

"We don't have a monopoly.
We have market share.
There's a difference."

– Steve Ballmer

4 Provider Competition

4.1 Introduction

Cloud services allow users to rent and use computer infrastructure for processing and storage over the Internet. A common way to pay for standardized cloud infrastructure instances, which offer a specific processing and/or storage capacity, is to obtain them on-demand, which means a client only pays for the time that the instance is used. Alternatively, as the previous chapter already described, providers may rent out the same instances as reserved instances. A reserved-only instance is an instance that is rented out to a client for a longer period of time, like a year, at a fixed price. In this case, the price does not depend on whether the client uses the instance or not. But then, providers might as well offer *two-part tariffs*, where a reduced fixed price is paid for a long timespan and an additional on-demand price applies for the actually utilized time. This can be reasonable because some cost factors (e.g. energy) depend on whether an instance is in use or not whereas other factors (e.g. building deprecation) occur in any case. A provider also might overbook resources that are currently not in use, which yields additional revenue and allows a lower reserved price. In exchange for the reserved price, the provider has to guarantee the availability of a reserved instance. An even more complex price model could also include a certain amount of use that is already covered by the fixed payment (comparable to inclusive minutes

in mobile phone contracts). In such *three-part tariffs*, only instance utilization that exceeds the included amount is charged the on-demand price (Figure 4.1).

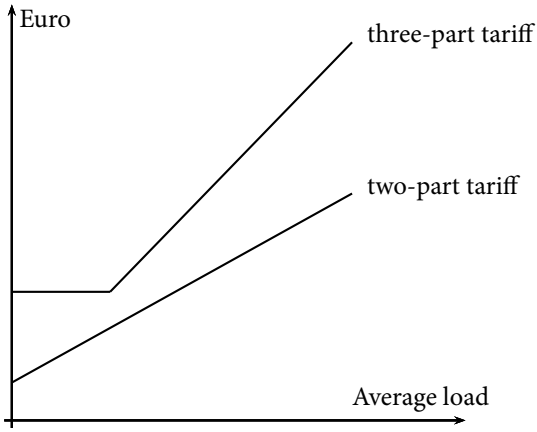


Figure 4.1: Two-part and three-part tariffs.

The previous chapter showed that with an on-demand price model, clients can have a financial benefit when they use a cloud service instead of building their own data center. Reserved instances reduce this client benefit. Different price combinations of two-part tariffs are differently attractive to clients depending on their load profiles. It is unclear, though, how this affects prices in a market in which several providers are present. Extending the previous results, we now investigate cloud instance pricing in the presence of provider competition.

In a scenario with competing offers for an identical service, rational clients are expected to choose the lowest price. Competition usually drives the price

lower than the monopoly price since every provider has an incentive to undercut the competitors' prices in order to attract their clients. But with a combined price model like two- or three-part tariffs, whether one offer is cheaper than another one depends on a client's load profile, particularly the average fraction of time over which a client uses its instances (*average load*). In order to determine the cheapest offer for them, clients have to know their average load; we assume precise knowledge about this value in the following. In such a setup, different combinations of reserved price, included resources and on-demand price might coexist.

This chapter explores a market where two providers maximize their profit from clients with different average load and offer an otherwise identical service at different price combinations. We assume that the providers know the distribution of average loads of the clients in the market. On that basis, we investigate how provider competition influences two- and three-part instance pricing. Our question is whether 'stable' constellations in such a market exist. In our understanding, stability requires that the current pricing schemes maximize both providers' profits given the competitor's pricing. This means that neither has an incentive to change its current pricing taking the competitor's behavior into account. Such a strategic interaction is analyzed here by means of game-theoretic models.

The chapter is organized as follows. Section 4.2 relates our contribution to other relevant literature. A game-theoretic approach to the problem is introduced in Section 4.3. Monopoly and duopoly market structures are then presented from a game-theoretic perspective and analyzed in the following sections: Section 4.4 deals with two-part tariffs, while Section 4.5 analyzes three-part tariffs. Asymmetric production costs of the two providers in a duopoly are investigated in Section 4.6. Section 4.7 concludes this chapter.

4.2 Related Work

Several studies in computer science explore provider competition in cloud computing with varying focus. Most closely related to the research in this chapter are those publications that also use a game-theoretic approach. Dynamics of service quality and pricing are covered in [31] and [80], where the existence of stable market shares has been shown for a duopoly and for n providers, respectively. We consider a duopoly where the providers offer the same service quality but use a more complex pricing scheme. An evolutionary approach on pricing cloud services is used in [9], which shows in simulations that an oligopoly market converges to a stable state over time. Instead, we make use of game theory for an analytical approach. [51] studies how cloud resource pricing is influenced by client competition, in contrast to our investigation of provider competition. The authors in [104] investigate a revenue-maximizing split-up of provider resources into reserved, on-demand and spot instances. This is similar to our research, where we determine the best on-demand and reserved price combination that is offered to the clients. However, provider competition is not covered in [104] and reserved and on-demand instances are separate offers that can be combined for different demand portions but do not converge in one product. While these pure on-demand and fixed price options can be considered as special cases of our combined price model (one price component is zero), we do not include an auction-based spot market.

In economics, combined pricing schemes are investigated independent of a cloud context. Since fundamental work in the early seventies [78], such two-part tariffs have been investigated with a focus on monopoly markets, but more recently also for duopoly markets. For monopolies, [90] provides a good survey. An interesting model for the duopoly case is proposed in [106] where customers have different preferences over the providers. A very similar model

is used in [44], where users differ in utilization rate and hence have a natural preference for one provider because of the prices. We use a slightly adapted model in this contribution and apply it to cloud instance pricing. There is also interesting research on three-part tariffs [19]; clients are not considered as a heterogenous group, though. We hence investigate three-part tariffs with regard to varying load among clients, which is more realistic in a cloud computing setting. We further contribute to the field by including two-part costs into the modeling and by briefly considering the option of multiple offers by each provider.

4.3 A Game-Theoretic Approach

4.3.1 Prices and Costs

We assume that a provider wants to maximize profit and each out of n clients wants to minimize cost. Profit and cost depend on the cloud service price. We now formalize two-part tariffs for cloud instances. A client i is characterized by the number of reserved instances x_i and a load factor a_i ($0 \leq a_i \leq 1$), which is the fraction of time that an instance is in use on average. The resulting price of one instance for one unit of time for a particular client, p_{instance}, is determined using the reserved price of an instance, p_{res}, and the price p_{od} that is charged on-demand for using the instance continuously for one unit of time (Equation 4.1).

$$p_{\text{instance}}\left(p_{\text{res}}, p_{\text{od}}, a_i\right) = p_{\text{res}} + a_i \cdot p_{\text{od}} \tag{4.1}$$

The total costs for a client are then the cost of one instance multiplied by the number of demanded instances x_i (Equation 4.2).

$$\begin{aligned} \text{client cost}\,(p_{\text{res}}, p_{\text{od}}, a_i, x_i) &= x_i \cdot p_{\text{instance}} \\ &= x_i \cdot (p_{\text{res}} + a_i \cdot p_{\text{od}}) \end{aligned} \tag{4.2}$$

We assume that the reservation of one instance causes fixed production cost c_{res} and the use of the instance causes additional variable operation cost c_{od} to the cloud provider (Equation 4.3). Such on-demand cost might be due to, e.g., an increased energy consumption of an instance under load.

$$c_{\text{instance}}\,(c_{\text{res}}, c_{\text{od}}, a_i) = c_{\text{res}} + a_i \cdot c_{\text{od}} \tag{4.3}$$

We further suppose there is a finite number of n clients. Provider profit is revenue minus production cost by all clients together (Equation 4.4).

$$\begin{aligned} &\text{provider profit}\,(p_{\text{res}}, p_{\text{od}}, c_{\text{res}}, c_{\text{od}}, a_1, \ldots, a_n, x_1, \ldots, x_n) \\ &\quad = \sum_{i=1}^{n} (x_i \cdot (p_{\text{instance}} - c_{\text{instance}})) \\ &\quad = \sum_{i=1}^{n} (x_i \cdot (p_{\text{res}} - c_{\text{res}} + a_i \cdot (p_{\text{od}} - c_{\text{od}}))) \end{aligned} \tag{4.4}$$

4.3.2 Provider Monopoly

In the presence of a single monopolistic cloud provider, a simple two-stage game can be used to determine the best reserved instance pricing from a provider perspective [62]. In a first move, the monopolistic provider determines a price and in a second move, the clients choose an amount of instances to use. The combined activity results in a specific situation (a specific price be-

ing paid for a specific amount of instances) with an according utility to both the provider and the clients. By anticipating the clients' move at each possible instance pricing, the provider can easily choose the price combination that maximizes its profit. This process is called *backwards induction*. When the profit that is generated by the expected service usage is described as a function over instance pricing, the cloud provider takes the clients' optimizing behavior into account. This leaves a simple decision problem to the provider, which only has to find the function's maximum for the n clients in the market in order to determine the best pricing. We denote the provider's monopoly profit from Equation 4.4 by $\text{profit}^M(p_{\text{res}}, p_{\text{od}})$. (We omit the dependencies on costs and client characteristics here and in the following for simplicity.)

4.3.3 Provider Duopoly

Now, we investigate a duopoly setting where two cloud providers offer the same service in a market with a finite number of clients. Both providers have identical production costs. Unlike the monopoly case, each cloud provider has to determine its optimal price not just with regard to the optimizing behavior of the clients but also with regard to the competitor's price choice. Therefore, each provider can expect to obtain only some fraction of the clients' demand, depending on the prices that are offered by both cloud providers. Since the expected demand accordingly does not only depend on the own instance pricing but also on the other provider's action, this decision problem can be modeled as a two-person non-cooperative game. The players are the two cloud providers A and B. The possible actions of the players are the possible price combinations at which they offer their service. We refer to such a combination of p_{res} and p_{od} of one provider as a *pricing*. For two-part tariffs, these pricings are $\left(p^A_{\text{res}}, p^A_{\text{od}}\right)$ and $\left(p^B_{\text{res}}, p^B_{\text{od}}\right)$. Both players offer their pricing at the same time. For each possible action profile (i.e. a combi-

nation of pricings), we determine the utility (expected profit) of each player. Contrary to the monopoly case, though, only the clients that prefer a particular provider have to be considered for this provider's utility. Therefore, the sum in Equation 4.4 only has to be taken over these clients. This depends on both offers and is further explained in Section 4.4.2. We denote the according duopoly profit of provider A and B as $\text{profit}^A(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ and $\text{profit}^B(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$. These are the utilities of the game (Figure 4.2).

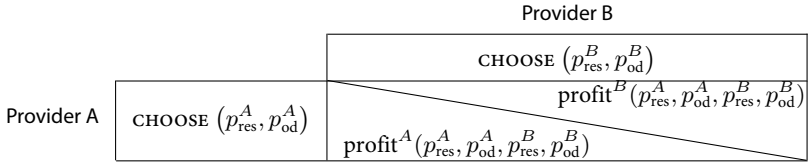


Figure 4.2: The non-cooperative duopoly game with two-part tariffs.

We examine three different pricing schemes in the following. Single two-part, multiple two-part and single three-part tariffs per provider are covered in separate sections (Table 4.1). Each section first presents the simpler monopoly case as a benchmark and then analyzes the game-theoretic model for the duopoly case.

Table 4.1: Overview of the different analysis cases.

# providers \ offers	1, two-part	k, two-part	1, three-part
1	Section 4.4.1	Section 4.4.3	Section 4.5.1
2	Section 4.4.2		Section 4.5.2

4.4 Two-Part Tariffs

4.4.1 Two-Part Tariffs in a Monopoly

We assume that any client could operate a server in its own data center at cost p_{server}. Under that assumption, a rational client is not willing to pay more than this amount for the number of cloud instances e that is necessary to match the server's performance. Any price combination where Condition 4.5 does not hold results in cloud cost that are higher than costs of an own data center to clients that have an individual average load of at least a_i.

$$\begin{aligned} e \cdot p_{\text{instance}} &\leq p_{\text{server}} \\ \Leftrightarrow e \cdot (p_{\text{res}} + a_i \cdot p_{\text{od}}) &\leq p_{\text{server}} \end{aligned} \tag{4.5}$$

Too high a pricing drives the affected clients out of the cloud. In case of a reserved-only pricing with $p_{\text{od}} = 0$, a client is willing to pay at most $p_{\text{res}} = \frac{p_{\text{server}}}{e}$ independent of its load characteristics. We observe that the highest possible pure reserved price $p_{\text{res}} = \frac{p_{\text{server}}}{e}$ is more profitable than any pricing with an on-demand price component $p_{\text{od}} > 0$ that meets Condition 4.5 ($p_{\text{res}} \leq \frac{p_{\text{server}}}{e} - a_{\max} \cdot p_{\text{od}}$ with $a_{\max} = \max(a_1, \ldots, a_n)$ as the maximum average load). This can be seen in Figure 4.3, which shows the instance price at a pure on-demand and at an example two-part tariff as well as the provision cost over the average load. The difference between cost and price is the profit of the provider and hence shall be maximized. The profit that is lost with a pricing that includes a positive on-demand price is indicated by the gray area in Figure 4.3.[1] There is a specific load at which both pricings produce the same instance price (straights intersect). Clients with a higher load than that do not meet Condition 4.5 and hence do not buy the service.

[1] Note that in our discrete model this is technically not an area as there are finitely many clients.

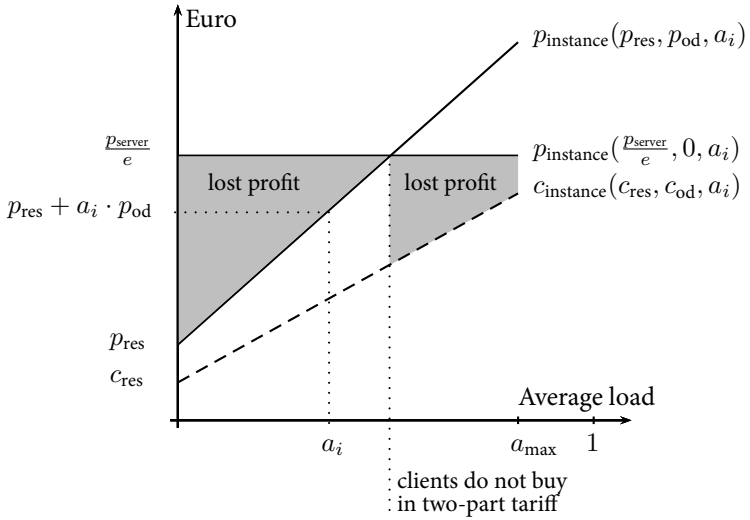


Figure 4.3: Monopoly pricing with two-part tariffs.

Formally, the monopoly profit from a reserved-only pricing is given by Equation 4.6. Including an on-demand price component p_{od} and adjusting the reserved price yields Equation 4.7. When we compare these two profits, we immediately observe that the reserved-only pricing (Equation 4.6) yields greater or equal profit compared to any other pricing (Equation 4.7) since Inequality 4.8 is always true.[2]

$$\text{profit}^M\left(\frac{p_{\text{server}}}{e}, 0\right) = \sum_{i=1}^{n}\left(x_i \cdot \left(\frac{p_{\text{server}}}{e} - c_{\text{res}} - a_i \cdot c_{\text{od}}\right)\right) \tag{4.6}$$

[2]Note that there is no possibility for a monopolistic cloud provider to generate a higher profit from not selling to the entire market since the formal argument also holds for any subgroup of clients with its specific maximal load.

$$\begin{aligned}&\text{profit}^M\left(\frac{p_{\text{server}}}{e} - a_{\max}\cdot p_{\text{od}}, p_{\text{od}}\right)\\ &\quad= \sum_{i=1}^{n}\left(x_i\cdot\left(\frac{p_{\text{server}}}{e} - a_{\max}\cdot p_{\text{od}} - c_{\text{res}} + a_i\cdot(p_{\text{od}} - c_{\text{od}})\right)\right)\end{aligned} \tag{4.7}$$

$$\begin{aligned}&\text{profit}^M\left(\frac{p_{\text{server}}}{e}, 0\right) - \text{profit}^M\left(\frac{p_{\text{server}}}{e} - a_{\max}\cdot p_{\text{od}}, p_{\text{od}}\right) \geq 0\\ &\qquad\qquad\Leftrightarrow \sum_{i=1}^{n} x_i\cdot a_i \leq \sum_{i=1}^{n} x_i\cdot a_{\max}\end{aligned} \tag{4.8}$$

Usually, there are clients with different average loads in the population, meaning that there exists an individual average load a_i that is strictly smaller than maximal average load $a_{\max}$. Thus, the above inequality is strict and hence the monopolistic cloud provider obtains strictly higher profits with a reserved-only pricing than with any other pricing including an on-demand price component.

4.4.2 Two-Part Tariffs in a Duopoly

For now, we assume that both providers offer a service at a single two-part pricing. Let us further assume a client population that features clients of varying average load a_i.[3] In the following we analyze the equilibrium behavior of the cloud providers.

Since each client tries to minimize its cost, it chooses the solution that is cheaper for its load factor a_i. For each action profile $(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ of the cloud providers A and B, the client population can be divided into three groups: In the first two groups the clients use the service from exactly one

[3] The abstract market model that is analyzed in [44] is very similar. We modify their solution to fit our setting of a finite number of clients and consider fixed and usage-based cost.

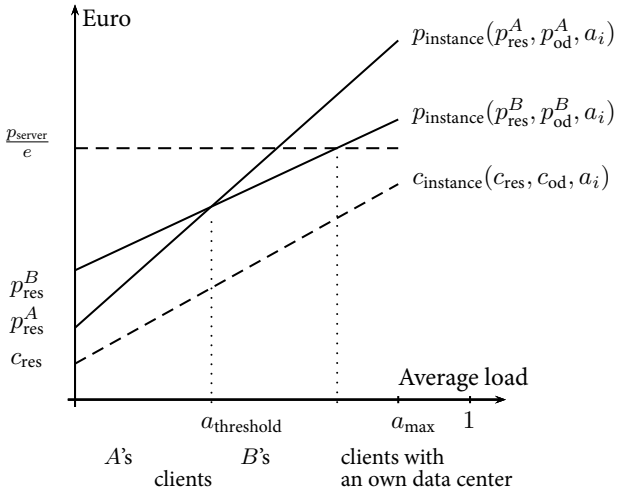


Figure 4.4: Duopoly pricing with two-part tariffs.

cloud provider (A or B) and in the third group they do not buy a service at all (because Condition 4.5 does not hold). Figure 4.4 illustrates this case. Formally, we distinguish the case of different on-demand prices and the special case where they are equal. First, if the on-demand prices are different ($p^A_{\text{od}} \neq p^B_{\text{od}}$), we can define a critical load $a_{\text{threshold}}$ where a client with this load is indifferent between buying from provider A or B (Equation 4.9).[4]

$$\begin{aligned} p^A_{\text{res}} + p^A_{\text{od}} \cdot a_{\text{threshold}} &= p^B_{\text{res}} + p^B_{\text{od}} \cdot a_{\text{threshold}} \\ \Leftrightarrow a_{\text{threshold}} &= \frac{p^B_{\text{res}} - p^A_{\text{res}}}{p^A_{\text{od}} - p^B_{\text{od}}} \end{aligned} \tag{4.9}$$

[4] The load a_i corresponds to the usage rate and $a_{\text{threshold}}$ is equivalent to the usage rate of the marginal consumer in [44].

For a threshold load $a_{\text{threshold}} > 0$, clients with an average load that is lower than $a_{\text{threshold}}$ prefer the offer with the lower reserved price. For instance, clients with $a_i < a_{\text{threshold}}$ prefer A's offer when $p^A_{\text{res}} < p^B_{\text{res}}$. Accordingly, if $a_{\text{threshold}} \in (0, 1)$, there are usually clients that prefer different offers and the market is divided. Note that it is also possible that $a_{\text{threshold}} > 1$ or $a_{\text{threshold}} < 0$, which means that the provider with the higher reserved price is not able to sell any service. Now consider the special case where both providers charge equal on-demand prices ($p^A_{\text{od}} = p^B_{\text{od}}$). If the reserved prices are also identical ($p^A_{\text{res}} = p^B_{\text{res}}$), the market is equally divided. Otherwise, if the reserved prices differ ($p^A_{\text{res}} \neq p^B_{\text{res}}$), the provider with the higher reserved price does not sell any service.

The profit of each provider is calculated based on the expected usage from the group of clients to which its service is sold. As the clients compare the offers from both providers, the pricing decision of one provider depends on the pricing decision of the other provider and therefore has to be taken into account. When provider A takes a specific offer $(p^B_{\text{res}}, p^B_{\text{od}})$ of provider B as given, A can determine profit$^A(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ by using Equation 4.4 while considering only those clients that prefer its service.

We now determine the *Nash equilibria* of the game. A Nash equilibrium is an action profile at which both providers mutually choose a best response (Section 2.2). A best response of a provider is its profit maximizing pricing at the current pricing of the competitor. Since in a Nash equilibrium both pricings are profit-maximizing, no provider has an incentive for a unilateral change of its pricing.

There is a unique Nash equilibrium. In this equilibrium, both providers offer identical two-part tariffs equal to marginal costs, $p^A_{\text{res}} = p^B_{\text{res}} = c_{\text{res}}$ and $p^A_{\text{od}} = p^B_{\text{od}} = c_{\text{od}}$. The argument is the following: If the competitor sets prices equal to marginal costs, then deviating with prices below marginal costs induces a loss and choosing prices above marginal costs implies not selling any-

thing. Therefore, no provider has an incentive to deviate, as long as the other provider offers prices equal to marginal costs. This shows the existence of the above equilibrium. To show its uniqueness, we now verify that there are no Nash equilibria with prices other than marginal costs.

Suppose provider B makes an offer non-equal to marginal costs: $(p^B_{\text{res}}, p^B_{\text{od}}) \neq (c_{\text{res}}, c_{\text{od}})$. Assume for simplicity that no client is running an own data center (Condition 4.5). Any client with an average load that meets Condition 4.10 prefers to use the service of provider A over B's service at the given pricings.

$$p^A_{\text{res}} + a_i \cdot p^A_{\text{od}} \leq p^B_{\text{res}} + a_i \cdot p^B_{\text{od}} \tag{4.10}$$

We distinguish three cases in order to show that B's pricing is not part of an equilibrium: first, where B makes positive profit; second, where B makes zero profit; and third, where the profit is negative. In these cases, either B itself has a more profitable option or A can attract some of B's clients and make positive profit and, subsequently, B has an incentive to deviate; hence there is no equilibrium with prices that differ from costs.

Consider two-part tariffs $(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ where B makes positive profit. When A offers the same pricing as B, the market is equally divided. When A chooses to offer the same on-demand price as B and slightly undercuts B's reserved price, then provider A is able to serve the entire market since all clients meet Condition 4.10 independent of their specific average load. Formally, provider A either undercuts provider B, yielding a profit as in Equation 4.11, or A chooses the same pricing as B, obtaining the profit in Equation 4.12. Comparing these two profits yields Equation 4.13.

$$\text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}\right) = \sum_{i=1}^{n}\left(x_i \cdot \left(p^A_{\text{res}} - c_{\text{res}} + a_i \cdot \left(p^A_{\text{od}} - c_{\text{od}}\right)\right)\right) \tag{4.11}$$

$$\text{profit}^A\left(p^B_{\text{res}}, p^B_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}\right) = \frac{1}{2}\sum_{i=1}^{n}\left(x_i \cdot \left(p^B_{\text{res}} - c_{\text{res}} + a_i \cdot \left(p^B_{\text{od}} - c_{\text{od}}\right)\right)\right) \tag{4.12}$$

$$\begin{aligned}&\text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}\right) - \text{profit}^A\left(p^B_{\text{res}}, p^B_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}\right)\\ &= \frac{1}{2}\text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}\right) + \frac{1}{2}\sum_{i=1}^{n}\left(x_i \cdot \left(p^A_{\text{res}} - p^B_{\text{res}} + a_i \cdot \left(p^A_{\text{od}} - p^B_{\text{od}}\right)\right)\right)\end{aligned} \tag{4.13}$$

This shows that as long as the profit from undercutting provider B (first term in Equation 4.13) is large enough to compensate for the price deduction (second term in Equation 4.13), which is chosen as small as possible, provider A slightly undercuts provider B's offer. When A undercuts B's pricing, then A serves the clients that were originally served plus B's former clients. A client that was previously attracted by A yields at least the same profit as before. If this would not be the case and the client would benefit from the new pricing, it would have chosen B's offer originally. Since B made positive profit, the new clients also yield positive profit for A. Accordingly, A cannot reduce its profit from undercutting B's pricing. Note that this is also true if A's profit was originally higher than (or equal to) B's profit. Also note that while undercutting B's reserved price is the better option compared to offering the same as B, it is not necessarily A's best response: A also has the option to set one price higher and the other price lower than B. Nevertheless, the argument is sufficient to show that an offer with at least one price over marginal cost is not an equilibrium price. If provider A hereby sets prices above marginal costs, the analog argument holds for provider B and therefore provider B has

an incentive to slightly undercut provider A's offer. Hence, as long as an offer makes positive profit, the best response of a provider is to make the same offer as the competitor but with a small discount to attract all clients.

Consider now two-part tariffs $(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ where B's overall revenue is equal to total instance production cost (zero profit). Provider A has no incentive to undercut the pricing of provider B in both reserved and on-demand price because this would lead to a negative profit and negative profits can always be avoided by setting prices equal to marginal cost. Since the average price of an instance for a client usually depends on its average load a_i, a pricing where $(p^B_{\text{res}}, p^B_{\text{od}}) \neq (c_{\text{res}}, c_{\text{od}})$ with zero overall profit means that some clients are profitable for provider B while others incur profit losses. A client with the load $a_i = \frac{c_{\text{res}} - p^B_{\text{res}}}{p^B_{\text{od}} - c_{\text{od}}}$ would generate exactly zero profits for provider B (Equation 4.14).

$$p^B_{\text{res}} - c_{\text{res}} + \frac{c_{\text{res}} - p^B_{\text{res}}}{p^B_{\text{od}} - c_{\text{od}}} \left(p^B_{\text{od}} - c_{\text{od}}\right) = 0 \tag{4.14}$$

Positive profits can be generated by clients with a higher (or lower) average load, depending on how provider B's pricing relates to its costs (Equation 4.15).

$$p^B_{\text{res}} - c_{\text{res}} + a_i \left(p^B_{\text{od}} - c_{\text{od}}\right) > 0 \tag{4.15}$$

$$\begin{cases} \text{for } a_i > \frac{c_{\text{res}} - p^B_{\text{res}}}{p^B_{\text{od}} - c_{\text{od}}} & \text{if } p^B_{\text{res}} < c_{\text{res}} \text{ and } p^B_{\text{od}} > c_{\text{od}} \\ \text{for } a_i < \frac{c_{\text{res}} - p^B_{\text{res}}}{p^B_{\text{od}} - c_{\text{od}}} & \text{if } p^B_{\text{res}} > c_{\text{res}} \text{ and } p^B_{\text{od}} < c_{\text{od}} \end{cases} \tag{4.16}$$

Thus, provider A can choose a pricing that attracts all profitable clients while all unprofitable clients still prefer the pricing provider B offers. Technically, the idea is to choose a reserved price and an on-demand price $(p^A_{\text{res}}, p^A_{\text{od}})$ so that the pricings of both providers and the costs have a common intersection

point. This pricing is derived by substituting the average load factor that leads to zero profit (for provider B) by the threshold average load (Equation 4.17).

$$\begin{aligned} 0 &= p^B_{\text{res}} - c_{\text{res}} + a_{\text{threshold}} \cdot \left(p^B_{\text{od}} - c_{\text{od}}\right) \\ &= p^B_{\text{res}} - c_{\text{res}} + \frac{p^B_{\text{res}} - p^A_{\text{res}}}{p^A_{\text{od}} - p^B_{\text{od}}} \cdot \left(p^B_{\text{od}} - c_{\text{od}}\right) \\ \Leftrightarrow p^A_{\text{od}} &= \frac{\left(p^B_{\text{res}} - p^A_{\text{res}}\right)\left(p^B_{\text{od}} - c_{\text{od}}\right)}{\left(c_{\text{res}} - p^B_{\text{res}}\right)} + p^B_{\text{od}} \end{aligned} \tag{4.17}$$

Provider A has an incentive to offer a reserved price $p^B_{\text{res}} < p^A_{\text{res}} < c_{\text{res}}$ (or $p^B_{\text{res}} > p^A_{\text{res}} > c_{\text{res}}$) and an on-demand price according to Equation 4.17. Figure 4.5 illustrates this. As a consequence, provider B keeps only clients that are unprofitable at its current pricing and therefore makes a negative profit. Hence, its pricing cannot be an equilibrium strategy and therefore no zero-profit equilibrium with pricings different from marginal costs exists.

Finally, consider the third case where B's profits are negative. In this case, provider B itself has an incentive to change its pricing to meet marginal costs and, accordingly, the pricing also cannot be an equilibrium strategy.

Summing up, no offer where the pricing does not equal production costs in both price components can be an equilibrium strategy and hence the equilibrium with $p^A_{\text{res}} = p^B_{\text{res}} = c_{\text{res}}$ and $p^A_{\text{od}} = p^B_{\text{od}} = c_{\text{od}}$ is unique. It is the only pricing that does not lead to any profit from any client and is hence the only pricing that does not offer an incentive to the competitor to aim for the profitable clients.[5]

This is very similar to Bertrand competition, a simple model for price competition where a good is offered by two competitors at a single price each and

[5] This is in accordance with the results in [44] where the equilibrium price equals the presumed zero cost. We note that our solution complements this by explicitly modeling reserved and on-demand costs. We emphasize that both price components have to be identical with the corresponding cost components to be in equilibrium.

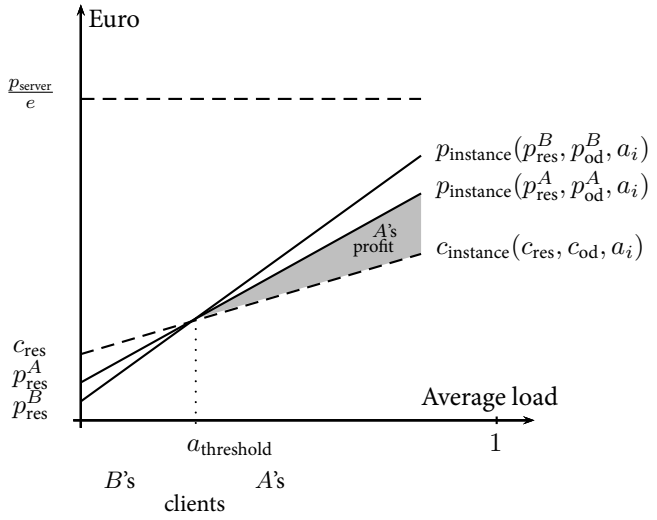


Figure 4.5: Zero-profit pricing is not necessarily an equilibrium strategy.

the only equilibrium is to make zero profits by setting prices equal to marginal costs [72]. Although this is a drastic indication and the limitations of the model have to be considered, it has to be noted that a two-part tariff as used in this model does not affect the general result of such a game setup. Is is important to notice, though, that neither different offers that are preferred by different clients nor the same offer where the reserved/on-demand price does not equal the reserved/on-demand cost can be an equilibrium strategy. Even if the overall profit is zero, there is always a subset of profitable clients in these cases.

4.4.3 Multiple Two-Part Tariffs per Provider

The results from the preceding section are not affected by several simultaneous offers of each provider. First, consider the monopoly case. When the provider makes several offers, each client chooses among them the offer that is the cheapest for this particular client. As a consequence, a client population with different average loads splits up among these different offers.[6] Any offer that is preferred by a client over another offer (because it lowers the client's cost) reduces provider profit accordingly. An additional offer can only increase the overall profit in comparison to a single offer when that single offer does not meet Condition 4.5 for all potential clients. Those clients would not participate in the market because they have too high an average load to accept the single offer. They can be attracted with an additional offer, though, when it has a lower on-demand price and hence meets Condition 4.5. Since the second offer also meets Condition 4.5 for all those clients that still prefer the first offer, the first offer then reduces provider profit and should be abandoned. But as mentioned above, Condition 4.5 holds for all potential clients anyway when the provider makes its most profitable single offer, which does not include an on-demand price. This means that a provider always reduces its profit by offering several price options.

Now, consider the duopoly case. As before for a single competitor offer, no profit that exceeds the profit of the competitor can be realized from any client. It is hence the best option to undercut a competitor offer in order to achieve a similar profit as the competitor from each client. Or, if not all clients are

[6] This is based upon the presumption that each client accepts the same offer for all of its demand. In the case that clients can split up their demand over several offers, then instead of the total demand of different clients, a different portion of every client's demand is best assigned on each offer. Accordingly, instead of x_i instances that are associated with a client's average load a_i, in a hybrid cloud scenario each load factor is associated with a total number of instances in the market that features this average load. Such a setting is described in [62] and it has the same main result.

profitable, to attract the profitable ones. Analog to the monopoly case, no additional offer can increase the provider's profit. It is not sufficient to respond to only one offer of the competitor when the competitor makes another offer that is still cheaper for some clients. Responding to all competitor offers, though, creates a situation where the demand of all profitable clients is met. In consequence, the more profitable option of a provider is to make the same number of offers as the competitor where each offer undercuts a counterpart or attracts profitable clients of the competitor. The only offer that does not provide the same incentive to the competitor again is a pricing that equals production costs. Thus, there is no equilibrium that features several distinct offers.

There are other reasons in favor of several offers, though. They can provide better knowledge about how much the instances are going to be actually used, for example, which can be of financial benefit since it eases overbooking of available capacities.

4.5 Three-Part Tariffs

4.5.1 Three-Part Tariffs in a Monopoly

Let us now consider that a reserved instance allows a certain amount of use without an on-demand charge. Accordingly, the tariff not only consists of a reserved and an on-demand price component, p_{res} and p_{od}, but also a third component r ($0 \leq r \leq 1$) that specifies the included usage, which is an average load over the reservation period (a fraction of time, analog to a_i) that is already covered with the reservation. When the actual average load of a client exceeds the included usage, the on-demand fee applies for the exceeding demand portion. The production costs remain two-part. Equation 4.18 states the provider profit for such a three-part pricing in a monopoly.

$$\text{profit}^M\left(p_{\text{res}}, p_{\text{od}}, r\right) = \sum_{i=1}^{n} \left(x_i \cdot \left(p_{\text{res}} - c_{\text{res}} + \max\left(a_i - r, 0\right) \cdot p_{\text{od}} - a_i \cdot c_{\text{od}}\right)\right) \tag{4.18}$$

The included amount of usage means that all clients that have an average load of $a_i \leq r$ have the same instance cost. For clients with an $a_i > r$, there is a linear increase of instance cost over load a_i. This is illustrated in Figure 4.6. Similar to the case of two-part pricing, a provider has no incentive to make an offer with a positive on-demand price in a monopoly. Compared to the highest possible reserved-only pricing, also with three-part tariffs an on-demand charge means that the provider has lower profit from clients that have a lower average load a_i than the client with the highest average load $a_{\max}$.

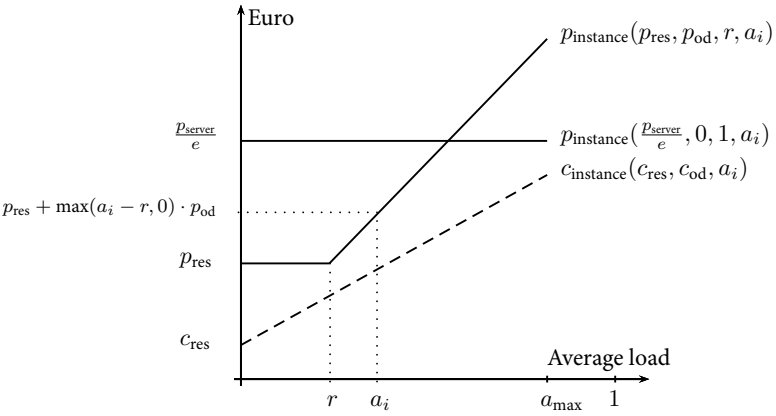


Figure 4.6: Monopoly pricing with three-part tariffs.

With three-part tariffs, a reserved-only pricing can be modeled either by including a sufficient amount of usage ($a_{\max} \leq r \leq 1$) or by setting the on-demand price to zero (Equation 4.19). When the included usage exceeds the highest average load ($r \geq a_{\max}$), the pricing may feature an arbitrary on-demand price since it is never charged. In Equation 4.20, the highest possible reserved price that attracts all clients again depends on a chosen on-demand price and is substituted accordingly. The reserved-only pricing (Equation 4.19) yields greater or equal profit compared to any other pricing (Equation 4.20) since Inequality 4.21 is true for all $r < a_{\max}$ and when clients differ in their average load.

$$\text{profit}^M\left(\frac{p_{\text{server}}}{e}, 0, 1\right) = \sum_{i=1}^{n}\left(x_i \cdot \left(\frac{p_{\text{server}}}{e} - c_{\text{res}} - a_i \cdot c_{\text{od}}\right)\right) \tag{4.19}$$

$$\begin{aligned}&\text{profit}^M\left(\frac{p_{\text{server}}}{e} - \max\left(a_{\max} - r, 0\right) \cdot p_{\text{od}}, p_{\text{od}}, r\right)\\ &= \textstyle\sum_{i=1}^{n}\left(x_i \cdot \left(\frac{p_{\text{server}}}{e} - \max\left(a_{\max} - r, 0\right) \cdot p_{\text{od}} - c_{\text{res}} + \max\left(a_i - r, 0\right) \cdot p_{\text{od}} - a_i \cdot c_{\text{od}}\right)\right)\end{aligned} \tag{4.20}$$

$$\sum_{i=1}^{n} x_i \cdot \max\left(a_{\max} - r, 0\right) \geq \sum_{i=1}^{n} x_i \cdot \max\left(a_i - r, 0\right) \tag{4.21}$$

4.5.2 Three-Part Tariffs in a Duopoly

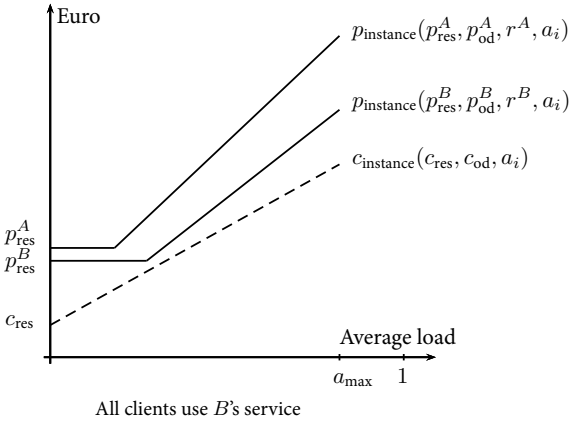
When two competing providers offer three-part tariffs, the result is similar to the outcome for two-part pricing, even if the analysis itself becomes more complex. A client i is indifferent between the offer from provider A and the offer from provider B if Equation 4.22 holds.

$$p_{\text{res}}^A + \max\left(a_{\text{i}} - r^A, 0\right) \cdot p_{\text{od}}^A = p_{\text{res}}^B + \max\left(a_{\text{i}} - r^B, 0\right) \cdot p_{\text{od}}^B \tag{4.22}$$

Contrary to the two-part tariff case, there can be no, one, two or infinitely many loads where a client is indifferent between the providers (Equation 4.22). As a consequence, with three-part tariffs there are now two different possibilities for a threshold load that separates the clients that prefer one offer over the other. First, like in the two-part case, different on-demand prices may lead to a threshold load $a^{\text{od}}_{\text{threshold}}$. Here, both providers charge on-demand prices since the included usage is exceeded. Second, an intersection might occur at a load where one provider offers a higher amount of included usage while the other provider already charges an on-demand price, which leads to a threshold load $a^{\text{r}}_{\text{threshold}}$. We compute both thresholds, $a^{\text{od}}_{\text{threshold}}$ and $a^{\text{r}}_{\text{threshold}}$, in Equation 4.23. Note that $a^{\text{od}}_{\text{threshold}}$ is required to be in the interval $[\max(r^A, r^B), 1]$.

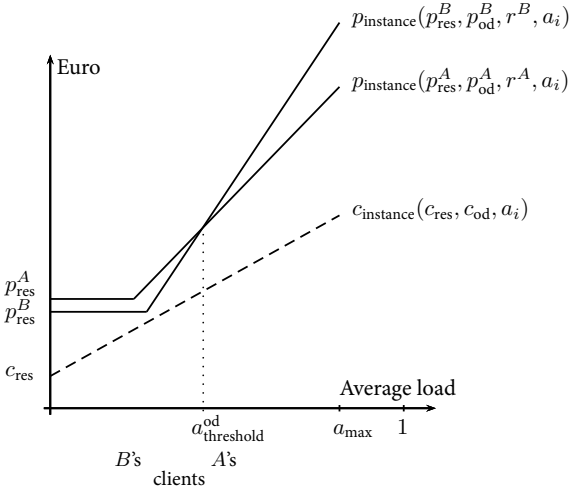
$$
\begin{aligned}
a^{\text{od}}_{\text{threshold}} &= \frac{p^B_{\text{res}} - p^B_{\text{od}} \cdot r^B - p^A_{\text{res}} + p^A_{\text{od}} \cdot r^A}{p^A_{\text{od}} - p^B_{\text{od}}} \quad \text{given} \quad p^A_{\text{od}} \neq p^B_{\text{od}} \\
a^{r}_{\text{threshold}} &= \begin{cases} \frac{p^A_{\text{res}} - p^B_{\text{res}}}{p^B_{\text{od}}} + r^B & \text{if} \quad r^A > r^B \quad \text{and} \quad p^A_{\text{res}} - p^B_{\text{res}} < p^A_{\text{od}} \cdot \left(r^A - r^B\right) \\ \frac{p^B_{\text{res}} - p^A_{\text{res}}}{p^A_{\text{od}}} + r^A & \text{if} \quad r^B > r^A \quad \text{and} \quad p^B_{\text{res}} - p^A_{\text{res}} < p^B_{\text{od}} \cdot \left(r^B - r^A\right) \end{cases}
\end{aligned}
\tag{4.23}
$$

Two pricings might intersect in none, both or either of these thresholds. Figures 4.7 and 4.8 illustrate this with some examples. In Example 1, provider B serves the whole market and in Examples 2 and 3, the two providers serve different market segments. Example 4 shows one possibility for a whole interval of load factors that satisfies Equation 4.22 (additionally an $a^{\text{od}}_{\text{threshold}}$ exists). Here, both offers have identical reserved prices and strictly positive included usage. Since all clients with a corresponding load are indifferent between both offers, we assume that they are equally divided between provider A and B. A second possibility to obtain such an interval is that the offers have identical on-demand prices ($p^A_{\text{od}} = p^B_{\text{od}}$) and the other price components fulfill $p^A_{\text{res}} - p^B_{\text{res}} = p^A_{\text{od}} \cdot (r^A - r^B)$.



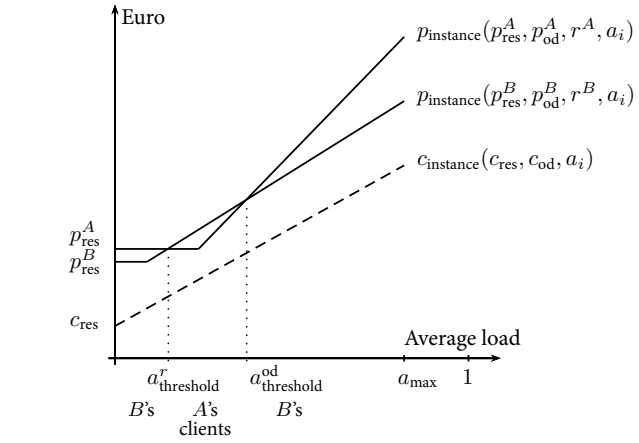
All clients use B's service

Example 1

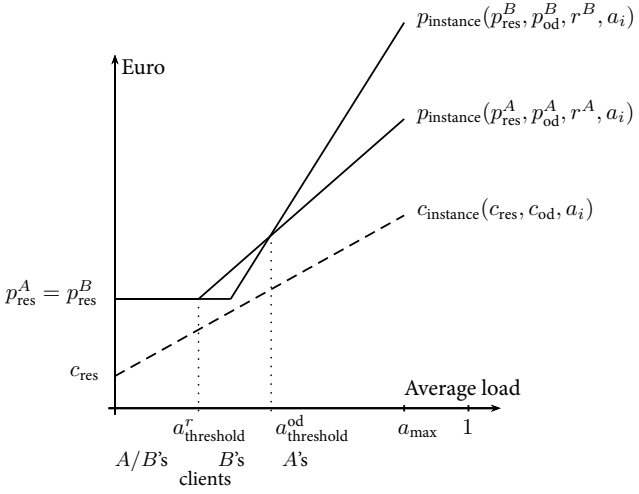


Example 2

Figure 4.7: Duopoly pricing with three-part tariffs.



Example 3



Example 4

Figure 4.8: Duopoly pricing with three-part tariffs (cont.).

Despite the greater freedom in pricing compared to two-part tariffs, there also exists only one unique Nash equilibrium. In this equilibrium, both providers offer identical three-part pricings equal to marginal costs, $p^A_{\text{res}} = p^B_{\text{res}} = c_{\text{res}}$, $p^A_{\text{od}} = p^B_{\text{od}} = c_{\text{od}}$ and no included usage $r^A = r^B = 0$. The argument for existence is analogous to the case of two-part tariffs. Deviating with prices below marginal costs induces a loss and choosing prices above marginal costs implies not selling anything. Including any usage free of charge also results in a loss. Therefore, no provider has an incentive to deviate, which shows the existence of the equilibrium. For its uniqueness, we now verify that there are no Nash equilibria with prices other than marginal cost. Similar to two-part tariffs, we assume there is another Nash equilibrium with $(p^B_{\text{res}}, p^B_{\text{od}}, r^B) \neq (c_{\text{res}}, c_{\text{od}}, 0)$. We distinguish three cases: first, where B makes positive profit; second, where B makes zero profit; and third, where the profit is negative. Either B itself has an incentive to deviate to a more profitable option or A can deviate from its offer in order to attract some of B's clients and make positive profit and, subsequently, B has an incentive to deviate. Accordingly, there is no such equilibrium.

Consider three-part tariffs $(p^A_{\text{res}}, p^A_{\text{od}}, r^A, p^B_{\text{res}}, p^B_{\text{od}}, r^B)$ where B makes positive profit. Provider A either undercuts provider B in at least the reserved price in order to serve the entire market (Equation 4.24) or A chooses the same pricing as B and subsequently serves half the market (Equation 4.25). Comparing these two profits yields Equation 4.26, which is positive when the price difference is sufficiently small. This implies that provider A has indeed an incentive to undercut provider B. With the same argument as in the two-part case, A always increases its profit from undercutting B's pricing.

$$\begin{aligned}&\text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, r^A, p^B_{\text{res}}, p^B_{\text{od}}, r^B\right)\\ &\quad= \sum_{i=1}^{n}\left(x_i \cdot \left(p^A_{\text{res}} - c^A_{\text{res}} + \max\left(a_i - r^A, 0\right) \cdot p^A_{\text{od}} - a_i \cdot c_{\text{od}}\right)\right)\end{aligned} \tag{4.24}$$

$$\begin{aligned}&\text{profit}^A\left(p^B_{\text{res}}, p^B_{\text{od}}, r^B, p^B_{\text{res}}, p^B_{\text{od}}, r^B\right)\\&\quad= \frac{1}{2}\sum_{i=1}^{n}\left(x_i \cdot \left(p^B_{\text{res}} - c^B_{\text{res}} + \max\left(a_i - r^B, 0\right)\cdot p^B_{\text{od}} - a_i \cdot c_{\text{od}}\right)\right)\end{aligned} \tag{4.25}$$

$$\begin{aligned}&\text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, r^A, p^B_{\text{res}}, p^B_{\text{od}}, r^B\right) - \text{profit}^A\left(p^B_{\text{res}}, p^B_{\text{od}}, r^B, p^B_{\text{res}}, p^B_{\text{od}}, r^B\right)\\&= \frac{1}{2}\cdot \text{profit}^A\left(p^A_{\text{res}}, p^A_{\text{od}}, r^A, p^B_{\text{res}}, p^B_{\text{od}}, r^B\right)\\&\quad+ \frac{1}{2}\sum_{i=1}^{n}\left(x_i \cdot \left(p^A_{\text{res}} - p^B_{\text{res}} + \max\left(a_i - r^A, 0\right)\cdot p^A_{\text{od}} - \max\left(a_i - r^B, 0\right)\cdot p^B_{\text{od}}\right)\right)\end{aligned} \tag{4.26}$$

Consider three-part tariffs $(p^A_{\text{res}}, p^A_{\text{od}}, r^A, p^B_{\text{res}}, p^B_{\text{od}}, r^B)$ where B makes zero profit. When $(p^B_{\text{res}}, p^B_{\text{od}}, r^B) \neq (c_{\text{res}}, c_{\text{od}}, 0)$, some clients are profitable while others are not and A has an incentive to attract the profitable ones. At a given pricing of B and given production costs, an appropriate pricing by A intersects B's pricing at the same average load as B's pricing intersects B's cost (B makes zero profit). In that case, the resulting thresholds not only separate clients that prefer A's or B's pricing, but also profitable and unprofitable clients. Unlike two-part tariffs we now may have two intersection points of provider B's pricing and its costs. At these intersection points clients would generate exactly zero profits for provider B (Equation 4.27). In order for B to make zero profit, at least one of them necessarily exists.

$$p^B_{\text{res}} - c_{\text{res}} + \max\left(a_i - r^B, 0\right)\cdot p^B_{\text{od}} - a_i \cdot c_{\text{od}} = 0 \tag{4.27}$$

Such a client has a load of either $a'_i = \frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}}$ or $a''_i = \frac{p^B_{\text{res}} - c_{\text{res}} - r^B p^B_{\text{od}}}{c_{\text{od}} - p^B_{\text{od}}}$. A client with load $a_i = r^B$ is profitable if and only if $\frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}} > r^B$, which means that $p^B_{\text{od}} < c_{\text{od}}$ is required for the existence of the on-demand intersection point (and vice versa). Then, clients with a lower (respectively higher) average load

than a_i'' are profitable. A reserved intersection point may exist only when a client with load $a_i = r^B$ is unprofitable. However, clients with a load $a_i < a_i'$ might not exist (in which case an on-demand intersection point exists). Note that profitable clients with a load $a_i < a_i'$ and profitable clients with a load $a_i > a_i''$ may coexist. Formally, provider B generates positive profits from the clients described by Inequality 4.28.[7]

$$p_{\text{res}}^B - c_{\text{res}} + \max\left(a_i - r^B, 0\right) \cdot p_{\text{od}}^B - a_i \cdot c_{\text{od}} > 0$$
$$\begin{cases} \text{for } a_i > \frac{p_{\text{res}}^B - c_{\text{res}} - r^B p_{\text{od}}^B}{c_{\text{od}} - p_{\text{od}}^B} \text{ and } \frac{p_{\text{res}}^B - c_{\text{res}}}{c_{\text{od}}} < r^B \text{ and } p_{\text{od}}^B > c_{\text{od}} \\ \text{for } a_i < \frac{p_{\text{res}}^B - c_{\text{res}} - r^B p_{\text{od}}^B}{c_{\text{od}} - p_{\text{od}}^B} \text{ and } \frac{p_{\text{res}}^B - c_{\text{res}}}{c_{\text{od}}} > r^B \text{ and } p_{\text{od}}^B < c_{\text{od}} \\ \text{for } a_i < \frac{p_{\text{res}}^B - c_{\text{res}}}{c_{\text{od}}} \text{ and } \frac{p_{\text{res}}^B - c_{\text{res}}}{c_{\text{od}}} \leq r^B \end{cases} \tag{4.28}$$

Provider A can choose a pricing that attracts at least a subset of these profitable clients, while all unprofitable clients still prefer the pricing of provider B. To determine an according pricing of A, we replace the average load factor in the profit function for one instance by the threshold average loads, $a_{\text{threshold}}^{\text{od}}$ and $a_{\text{threshold}}^{r}$. For instance, for a chosen on-demand price and a chosen included usage, the reserved price that creates an $a_{\text{threshold}}^{od}$ calculates according to Equation 4.29 (Figure 4.9, Example 5). When instead the on-demand and the reserved price are chosen, the included usage that creates an $a_{\text{threshold}}^{\text{r}}$ then calculates as presented in Equation 4.30 (Figure 4.9, Example 6).

[7] If $p_{\text{od}}^B > c_{\text{od}}$ and $\frac{p_{\text{res}}^B - c_{\text{res}}}{c_{\text{od}}} > r^B$, then provider B makes positive profit with all clients (a client with load $a_i = r^B$ is profitable and no on-demand intersection point exists). This contradicts the assumption that provider B makes zero profit.

$$
\begin{aligned}
0 &= p^B_{\text{res}} - c_{\text{res}} + \max\left(a^{\text{od}}_{\text{threshold}} - r^B, 0\right) \cdot p^B_{\text{od}} - a^{\text{od}}_{\text{threshold}} \cdot c_{\text{od}} \\
\Leftrightarrow p^A_{\text{res}} &= p^B_{\text{res}} + r^A \cdot p^A_{\text{od}} - r^B \cdot p^B_{\text{od}} + \frac{\left(p^B_{\text{res}} - c_{\text{res}} - r^B \cdot p^B_{\text{od}}\right) \cdot \left(p^A_{\text{od}} - p^B_{\text{od}}\right)}{\left(p^B_{\text{od}} - c_{\text{od}}\right)}
\end{aligned}
\tag{4.29}
$$

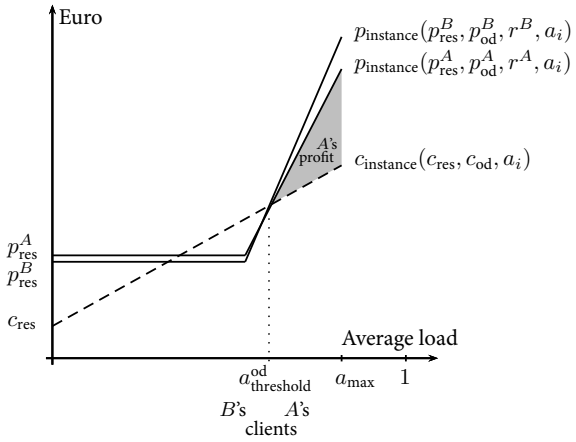
$$
\begin{aligned}
0 &= p^B_{\text{res}} - c_{\text{res}} + \max\left(a^{\text{r}}_{\text{threshold}} - r^B, 0\right) \cdot p^B_{\text{od}} - a^{\text{r}}_{\text{threshold}} \cdot c_{\text{od}} \\
\Leftrightarrow r^A &= \frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}} - \frac{p^B_{\text{res}} - p^A_{\text{res}}}{p^A_{\text{od}}}
\end{aligned}
\tag{4.30}
$$

At least one of the following three possibilities yields strictly positive profit for provider A by attracting profitable clients from Inequality 4.28.

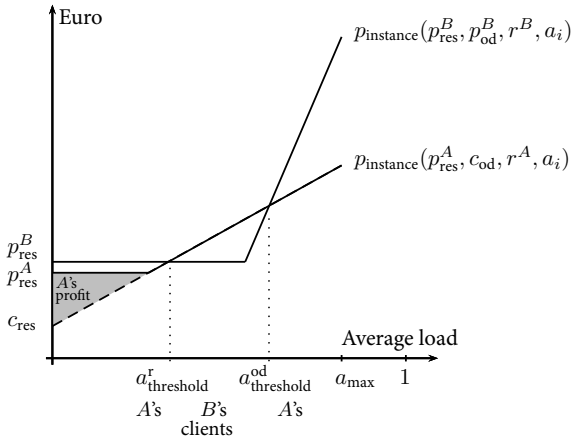
If $p^B_{\text{od}} > c_{\text{od}}$ and $\frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}} < r^B$: The on-demand price of B is undercut while the amount of included usage is set the same as B's $\left(p^B_{\text{od}} > p^A_{\text{od}} > c_{\text{od}}, r^A = r^B\right)$. Then the reserved price is determined using Equation 4.29 (Figure 4.9, Example 5).

If $p^B_{\text{od}} < c_{\text{od}}$ and $\frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}} > r^B$: The on-demand price of B is overcut while the amount of included usage is set the same as B's $\left(p^B_{\text{od}} < p^A_{\text{od}} < c_{\text{od}}, r^A = r^B\right)$. Then the reserved price is determined using Equation 4.29.

If $\frac{p^B_{\text{res}} - c_{\text{res}}}{c_{\text{od}}} < r^B$: The reserved price of B is undercut while the on-demand price is set at marginal cost $\left(p^B_{\text{res}} > p^A_{\text{res}} > c_{\text{res}}, p^A_{\text{od}} = c_{\text{od}}\right)$. Then the according amount of included usage is determined using Equation 4.30 (Figure 4.9, Example 6).



Example 5



Example 6

Figure 4.9: Provider A takes profitable clients with low average load.

Note that while this is sufficient to show that B's zero-profit pricing cannot be an equilibrium strategy when $(p^B_{\text{res}}, p^B_{\text{od}}, r^B) \neq (c_{\text{res}}, c_{\text{od}}, 0)$, the three possibilities do not necessarily maximize A's profit.[8]

Finally, if B's profits are negative, provider B itself has an incentive to deviate to marginal cost pricing.

In consequence, the equilibrium with $p^A_{\text{res}} = p^B_{\text{res}} = c_{\text{res}}$, $p^A_{\text{od}} = p^B_{\text{od}} = c_{\text{od}}$ and $r^A = r^B = 0$ is unique and analogous to the two-part tariff case. While the included usage in the tariff structure changes how a market can be split up between providers, a positive amount of included utilization time is not part of an equilibrium pricing.

4.6 Two-Part Tariffs with Asymmetric Production Costs

We now investigate two-part tariffs for the case where the production costs of both providers are different, $\left(c^A_{\text{res}}, c^A_{\text{od}}\right) \neq \left(c^B_{\text{res}}, c^B_{\text{od}}\right)$. Suppose provider A offers a different on-demand and reserved price than B. One of the following two cases always applies since the roles of A and B are interchangeable.

First, if one or both cost components of provider A are smaller than the corresponding price component of B, then A can attract the whole market and obtain a positive profit with prices between A's costs and B's prices.[9] Such a pricing $\left(p^A_{\text{res}}, p^A_{\text{od}}\right)$ of provider A is given in Inequality 4.31.

$$\begin{aligned} c^A_{\text{res}} < p^A_{\text{res}} < p^B_{\text{res}},\ c^A_{\text{od}} \leq p^A_{\text{od}} \leq p^B_{\text{od}} \quad &\text{if } c^A_{\text{res}} < p^B_{\text{res}} \text{ and } c^A_{\text{od}} \leq p^B_{\text{od}} \\ c^A_{\text{res}} \leq p^A_{\text{res}} \leq p^B_{\text{res}},\ c^A_{\text{od}} < p^A_{\text{od}} < p^B_{\text{od}} \quad &\text{if } c^A_{\text{res}} \leq p^B_{\text{res}} \text{ and } c^A_{\text{od}} < p^B_{\text{od}} \end{aligned} \tag{4.31}$$

[8] This can be easily observed, e.g. in Example 5, where the first possibility attracts all profitable clients of B, but only those with $a_i < a^r_{\text{threshold}}$ are profitable at A's pricing, and it would be better to choose a higher on-demand price that does not attract as many clients but those who are attracted yield some profit.

[9] The argument is similar to the case in symmetric costs where provider B makes positive profit.

Example 7 in Figure 4.10 illustrates this for the case that B offers its costs.

Second, suppose that one cost component of A is higher and the other cost component is lower than the corresponding price components of B. In order to determine an offer that attracts all profitable clients, A can choose its pricing so that it has a common intersection point with A's costs and B's pricing (Equation 4.32). Like this, the threshold load that separates clients that prefer A's offer over B's offer also separates the clients that are profitable to A from those that are not.[10] We obtain such an intersection point by substituting the load where B's pricing and A's costs are identical by threshold load $a_{\text{threshold}}$, which is the intersection of both pricings (Equation 4.32).

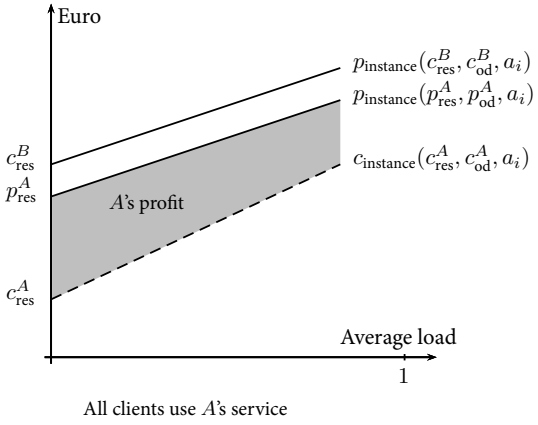
$$\begin{aligned} 0 &= p^B_{\text{res}} - c^A_{\text{res}} + a_{\text{threshold}} \cdot \left(p^B_{\text{od}} - c^A_{\text{od}}\right) \\ &= p^B_{\text{res}} - c^A_{\text{res}} + \frac{p^B_{\text{res}} - p^A_{\text{res}}}{p^A_{\text{od}} - p^B_{\text{od}}} \cdot \left(p^B_{\text{od}} - c^A_{\text{od}}\right) \\ \Leftrightarrow p^A_{\text{od}} &= p^B_{\text{od}} - \frac{\left(p^B_{\text{res}} - p^A_{\text{res}}\right) \cdot \left(p^B_{\text{od}} - c^A_{\text{od}}\right)}{p^B_{\text{res}} - c^A_{\text{res}}} \end{aligned} \tag{4.32}$$

If provider A offers a reserved price between A's reserved cost and B's reserved price (Inequality 4.33) and determines its on-demand price using Equation 4.32, then A attracts exactly those clients that yield positive profit for A whereas the remaining clients continue to buy their service from B.

$$\begin{aligned} c^A_{\text{res}} < p^A_{\text{res}} < p^B_{\text{res}} \quad &\text{if } c^A_{\text{od}} > p^B_{\text{od}} \text{ and } c^A_{\text{res}} < p^B_{\text{res}} \\ c^A_{\text{res}} > p^A_{\text{res}} > p^B_{\text{res}} \quad &\text{if } c^A_{\text{od}} < p^B_{\text{od}} \text{ and } c^A_{\text{res}} > p^B_{\text{res}} \end{aligned} \tag{4.33}$$

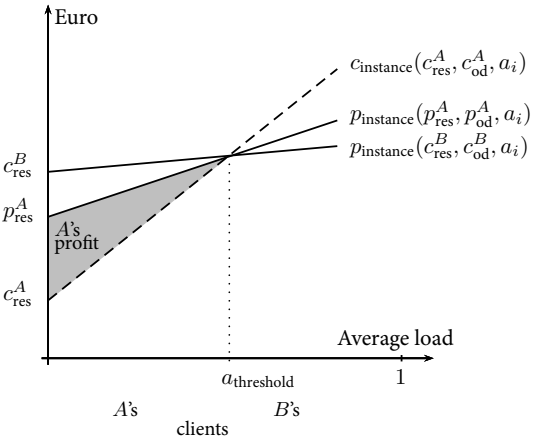
Figure 4.10, Example 8 illustrates this for the case that B offers its costs.

[10] The argument is similar to the case in symmetric costs where provider B makes zero profit but its prices do not equal production costs. Here we use A's specific costs.



All clients use A's service

Example 7



Example 8

Figure 4.10: When B offers its costs, A can make positive profit.

In consequence, no Nash equilibria exist with different prices. Since the above arguments hold for $\left(p^A_{\text{res}}, p^A_{\text{od}}\right) = \left(c^A_{\text{res}}, c^A_{\text{od}}\right)$ and $\left(p^B_{\text{res}}, p^B_{\text{od}}\right) = \left(p^B_{\text{res}}, p^B_{\text{od}}\right) = \left(c^B_{\text{res}}, c^B_{\text{od}}\right)$, and in contrast to symmetric costs, offering pricings equal to marginal costs is not a Nash equilibrium with asymmetric costs.

The above cases show that provider A has an incentive to attract clients from B that are profitable for A. Again, consider the (second) case where one cost component is higher and the other cost component is lower than the corresponding cost of the competitor. Suppose that B makes an offer and A takes the profitable clients. Then the providers offer a similar pricing but only serve clients that feature a load at which the providers have a lower production cost than their competitor. No provider has an incentive to attract clients from their competitor since these clients would be unprofitable to them at an offer that is attractive to the clients. Instead, both providers have an incentive to increase revenue from their own clients only. When they make the same offer as the competitor, though, they loose half their clients to the competitor (and also attract half the unprofitable competitor clients). Deviating to an offer with prices between the current prices as described above, though, increases revenue while the market shares are maintained. The closer the pricing is to the competitor's offer, the higher are revenue and profit. Since it is (theoretically) always possible to find an offer that is even closer to (but never the same as) the competitor's offer, there is no best response to such an offer and hence no Nash equilibrium exists.

The nonexistence of Nash equilibria is caused by a mathematical singularity due to the assumption of an equally divided market in case of identical offers. In reality, such a singularity probably does not exist. The smaller the price difference, the more important other factors may be, which are not explicitly considered in our model. It is also practically impossible to differentiate prices by arbitrarily small amounts. Accordingly, the result that no equilibria exist

is a rather theoretic result. The fact that the providers try to push their offers towards their competitor's offer nevertheless is an interesting observation.

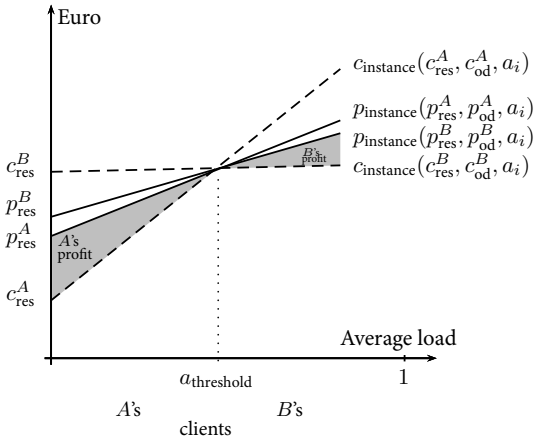
A slight modification of the equilibrium concept allows us to recover the existence of equilibria. We use ε-equilibria to show that there are conditions where no provider has a *significant* incentive to deviate, which seems to be a realistic candidate for a stable market situation. In contrast to usual Nash equilibria, an outcome is an ε-equilibrium when neither party can increase its utility by ε or more by deviating to another strategy. We have shown that both providers benefit only from deviating to an offer that is closer to the competitor's offer: If A deviates, the profit from any offer that is obtained using Equation 4.32 and that has on-demand and reserved prices between A's and B's offer is higher than the profit from A's original offer. The market shares are maintained. The profit at the competitor's offer from the same market share hence represents an upper bound. An action profile $(p^A_{\text{res}}, p^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}})$ is an ε-Nash equilibrium when the difference between the current profit and the upper bound for possible profits is smaller than ε. This means that provider A is unable to sufficiently increase its profit at the current offer by B and vice versa. More precisely, the pricings $(\bar{p}^A_{\text{res}}, \bar{p}^A_{\text{od}}, \bar{p}^B_{\text{res}}, \bar{p}^B_{\text{od}})$ constitute a ε-Nash equilibrium if Equation 4.34 applies.

$$\begin{aligned}
&\text{profit}^A(\bar{p}^A_{\text{res}}, \bar{p}^A_{\text{od}}, \bar{p}^B_{\text{res}}, \bar{p}^B_{\text{od}}) - \text{profit}^A(p^A_{\text{res}}, p^A_{\text{od}}, \bar{p}^B_{\text{res}}, \bar{p}^B_{\text{od}}) < \varepsilon \quad \text{for all } (p^A_{\text{res}}, p^A_{\text{od}}) \\
&\text{profit}^B(\bar{p}^A_{\text{res}}, \bar{p}^A_{\text{od}}, \bar{p}^B_{\text{res}}, \bar{p}^B_{\text{od}}) - \text{profit}^B(\bar{p}^A_{\text{res}}, \bar{p}^A_{\text{od}}, p^B_{\text{res}}, p^B_{\text{od}}) < \varepsilon \quad \text{for all } (p^B_{\text{res}}, p^B_{\text{od}})
\end{aligned} \tag{4.34}$$

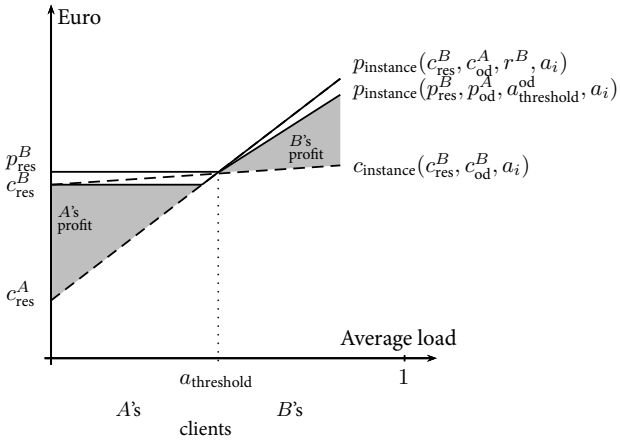
Accordingly, while there are no regular Nash equilibria when production costs of the two providers are asymmetric, there is an infinite amount of ε-equilibria. In these equilibria, usually both providers have positive profit. In the marginal cases where one provider offers its production costs, this provider has zero

profit while the other one makes positive profit. We observe that the individual profit that a provider can yield from its market share at an equilibrium pricing can be as high as the cost difference between both providers. However, the total profit of the two providers is limited in all equilibria because this maximum cannot be reached by both providers at the same time: When one provider maximizes its profit from clients for which the provider features lower production cost, its pricing also creates an upper bound for the competitor's pricing (Figure 4.11, Example 9). With three-part tariffs, this is not necessarily the case. For example, suppose that provider B wants to attract clients with a higher average load than $a^{\text{od}}_{\text{threshold}}$. Instead of a two-part offer, B may offer an included usage up to the threshold load. If the reserved price is chosen accordingly, then B itself has the same utility as in the two-part equilibrium pricing since the profits from its market share remain the same (Figure 4.11, Example 10). Nevertheless, this usually allows a higher pricing for A. As long as A's prices do not exceed B's costs, B has no incentive to deviate. Similarly, A might prefer to include a significant amount of usage at a profitable reserved price and combine it with a high on-demand price, which is not charged from A's clients anyway since their average load is too small. This raises the upper bound for B's pricing in B's market share. Although A usually is not able to utilize the full potential profit (difference in costs) like this, depending on B's pricing this is also true for two-part offers in most cases and, accordingly, such a three-part can be the better option.

While we refrain from calculating all equilibria in three-part tariffs with asymmetric costs or from providing formal proof for their existence, we note that such equilibria very likely exist. However, three-part tariffs provide an interesting potential for cooperation since the providers can significantly increase their profits in comparison to two-part offers (Figure 4.11).



Example 9



Example 10

Figure 4.11: Three-part tariffs can increase the equilibrium profit.

4.7 Conclusion and Implications on the IaaS Market

We showed that in a duopoly market for cloud infrastructure services, neither two-part or three-part offers result in equilibria where symmetric providers have positive profit. For asymmetric production costs, on the other hand, there exists an infinite number of ε-equilibria with two-part tariffs in which usually both providers have positive profit. Three-part tariffs may increase this profit for both providers.

However, there are some limitations of our market model that may be addressed in future research. For instance, we assume that the clients only differ in the average utilization rate (and the number) of reserved instances. Since clients do not participate in the market if their cost for own infrastructure is lower than the cost of cloud infrastructure, an upper price bound exists. This upper bound is based on the per-server cost of a data center and is assumed to be constant for all clients. At a given pricing, the average instance price of different clients depends on their average load via the on-demand price. In consequence, an offer with an on-demand price of zero is lower or higher than data center cost for all clients at the same time. While most data center cost factors can be regarded similar for different clients, building a data center is a huge investment. So is paying a reservation fee for a long time in advance. Depending on the financial standing of a client, some may not be able to afford such an investment. In addition, land, buildings or part of the infrastructure might already exist in some cases. This affects the cost of data center capacity. While differentiated costs for an own data center influence the monopoly pricing, the equilibrium pricing in a duopoly does not change because it is linked to the production costs and is therefore usually lower than the hypothetical data center cost anyway. Some clients, on the other hand, might not be able to raise enough money to invest in an own data center or have a low planning reliability and cannot be certain about their future amount of infrastructure

use. In these cases, pure on-demand services may be the only option and the providers might want to make such an offer in order to meet the demand of such clients. Clients with a low average load (in comparison to their peak demand) pay a lot less in a tariff without a reserved price. In consequence, the provider can be expected to ask a very high on-demand price in order to avoid that many of these clients also deviate to the pure on-demand tariff.

Although the model not only assumes symmetric providers but also considers different production costs, it does not regard limits of available capacity or marketing factors other than pricing. These model constraints also exist in classic Bertrand competition and it is obvious that the unaccounted factors are important in reality. What can be derived from the model is the fact that competition in a duopoly is potentially sufficient to keep prices low. This does not change with two-part or three-part tariffs for reserved cloud instances. Since no provider makes positive profit at the equilibrium pricing, it is likely that they take actions to reduce the competitive pressure. In this regard, asymmetries of the providers become important, which can lead to strictly positive profit of one or both providers as was shown in Section 4.6. Additionally, asymmetries might enable a provider to strive for a monopoly position by destructive competition or constitute a significant market entry barrier. Possible asymmetries other than different production costs are e.g. available resources or service quality. The providers accordingly have an incentive to differentiate themselves from their competitor in this regard.

An option to make profit without the need of asymmetries is that the providers refuse to compete and mutually agree to maintain arranged market shares on a very high price level with accordingly high profits. This is especially likely to happen when capacities are limited and the demand in the market is larger than what the providers could possibly meet altogether. The providers have an incentive to develop their capacities as long as it generates enough profit to make up for the investment. When the available capacity

exceeds overall demand, the providers compete for market shares. In order to split the market, providers can target different customer groups with their marketing. The combined price model for cloud instances offers an easy way to split up the market among several providers based on the clients' average utilization time of reserved instances. A provider just has to combine a higher on-demand price component with a lower reserved price component compared to a competitor in order to attract a different group of clients. Such mutually profitable pricing agreements between providers may raise the opinion that oligopolies require some form of regulation. Especially on a global and fast-developing market like the IaaS market, it appears challenging to reveal such collusive behavior. It seems to be an even greater challenge to find a common jurisdiction to enforce market regulation on a global level.

"Location, location, location."

– Real estate aphorism

5 Data Centers for Processing and Storage in Separate Locations

5.1 Introduction

All cloud services are eventually based on processing and storage. These can in turn be obtained in form of services and such hardware-bound services are referred to as *Infrastructure-as-a-Service (IaaS)* (Chapter 2). In addition to an abstraction of the actual hardware that is running underneath, transparency of IaaS also means an undetermined location of this hardware. IaaS providers are free to place data centers at any place with network access.

Clustering the provision of services is interesting for providers due to economies of scale. Local conditions like cheap power or a cool climate can lower operating costs even further. In practice, these possibilities are limited by technical restrictions, risk awareness and law.

Restrictions and savings potential are not necessarily the same for all service types. This chapter explores the possibility of separate processing and storage centers and their ability to compete with centers that combine those resources in one location. It takes into account that storage and processing, though different service types, affect each other: Both might handle the same data. Market dynamics in our model are not determined by different service qualities

as in related work (Section 5.2) but by scale and location of provider facilities. This work contributes a new perspective on the future market for cloud infrastructure and its geographical development. The question is whether and under what conditions several facility types can coexist in a stable market situation.

Considerations for storage and data center placement regarding their relative as well as their geographical location are discussed in Section 5.3. A game-theoretic market model that combines these factors is given and analyzed throughout Section 5.4. This model is extended in Section 5.5 in order to better consider client diversity. Section 5.6 states implications on the actual cloud that can be derived from these theoretic observations. A discussion of further aspects of the model and perspectives for future work are presented in Section 5.7.

5.2 Related Work

Cloud provider competition is the subject of some game-theoretic work regarding service quality and pricing. The existence of stable market shares in a duopoly [31] and recently also for n competitors [80] has already been shown. This chapter proposes a model for different but dependent service markets (different service types instead of service qualities) and analyses stable states in this set of markets.

Optimal placement of data centers is extensively discussed in [5]. Climate as a factor is specifically addressed in [38], but there seems to be a lack of scientific material that evaluates the effects of climate on data center economics. We discuss possible economies of location with a focus on their different impact on storage and processing facilities and provide an analytical perspective

on the question whether separately located facilities can exist in a stable market situation.

When focusing on data center location, data protection directives are important as storage of personal data might be regulated. The European data protection supervisor talks about the role of cloud providers and EU law implications [50]; US law is discussed in [42]. Apart from legal reasons, widely discussed privacy and security concerns (e.g. [55, 82]) might make customers more sensitive to storage location. While these factors can motivate a separation of storage and processing, they are hard to assess. Our model explores the existence of stable markets with separate facilities with a focus on economic factors.

Effects of cloud virtualization and remote data access on I/O performance are explored in [93, 14]. These practical findings are important when storage and processing are separated in different services and locations as is discussed in this chapter.

5.3 Placing Storage and Processing Infrastructure Sites

5.3.1 Separating Storage and Processing as Products

Local separation of storage and processing might appear impractical at first glance: Both services are associated with each other as processing generally involves data. While separate storage services make sense for archival purposes, exclusive processing usually cannot be utilized on its own. Combining both resources in one product thus appears to be a more sensible choice. Accordingly, processing usually is provided together with a certain amount of instance storage in today's infrastructure cloud market. Stand-alone storage is common practice, though.

Whenever data has to be shared between several processing instances, using instance storage is problematic as it is inaccessible from other instances. When instances are booted and shut down to flexibly adapt to actual processing demand, a lot of data management becomes necessary as the temporary instance storage is abandoned together with the instance. A separate shared storage like a distributed file system on block storage instances is far more handy. It can be accessed by independent processing instances which do not have to provide any disk storage. Such a setup is a lot more flexible for clients, who can scale storage and processing independently and also can combine services of different providers. It thus makes sense to provide storage and processing resources in separate products. Providers gain the possibility of separate facilities for resource types and can specialize on just storage or processing services.

Separating processing and storage in different products does not imply that corresponding hardware is placed in different locations. As a lot of traffic between the services can be expected, latency and traffic cost rather suggest to keep both resources close together. Providing both resources from the same facility can offer performance similar to that of instance storage and does not cause Internet traffic. There are some reasons in favor of a separation of both resources in different locations, though.

5.3.2 Separating Storage and Processing Locations

Most data center operating costs are caused by administration and energy. Automatization can reduce the average administration cost in larger data centers, which usually also have a better power usage effectiveness. Energy cost is not only affected by size, but also a lot by a data center's location. From a worldwide perspective, energy prices vary a lot. Cooler climate in some areas allows free-air cooling, which keeps both energy consumption and investments

in cooling equipment down. From an economic point of view, combining economies of scale and a locational advantage by operating huge data centers in cool areas with cheap power supply appears to be the only sensible choice. On the other hand, this may not be a good idea due to the following reasons.

Loss of data can be considered a lot worse than failure of processing as the latter should only be a temporary effect in most cases. As a consequence, safety from natural disasters might have more weight than e.g. climate during the selection of storage center locations. By building two separate facilities, both can gain from better location.

Regulation of private data is another issue that can drive storage and processing facilities apart. Imposed by European privacy law, such data has to be kept on European territory or areas of comparable protection [50]. These legal boundaries fragment the Internet in several zones that limit the technical freedom of storage deployment. Personal data might be processed in other zones, though, in an anonymized or pseudonymized form.

Data that is stored in the cloud is beyond clients' control as internal activities of the provider are hidden. Data recovery is doubtful in case that the service shuts down e.g. due to legal issues or bankruptcy. It also might be deleted in case a client cannot pay for the service. In consequence, clients may refrain from cloud storage options and keep vital data in their own storage facilities while they benefit from cheap and flexible cloud processing services at the same time.

5.4 Game-Theoretic Model

5.4.1 Setup

A simple evolutionary game-theoretic model (Section 2.2) is hereby proposed to identify stable market shares of separate facilities for storage and process-

ing services. Required conditions are determined regarding economic factors. Risk and law are considered later in Section 5.6.2.

The model distinguishes the two service types *storage* and *processing* and the three different facility strategies p (process), s (store) and c (combine). While c means operation of storage and processing in one facility, the strategies p and s stand for an exclusive operation of one service type in the facility. Any parameter or function that is defined specifically for a service type is indexed accordingly while facility strategies are specified as a function parameter. (This has to be distinguished because e.g. cost for storage may differs between the storage-only and the combined facility.) Variable x denotes the market ($x \in$ processing, storage) and y denotes a facility strategy ($y \in p, s, c$) when indicated.

An IaaS provider has to choose a facility strategy and passes on data center operation and investment costs to the clients. Constant R_x stands for reference amortization costs of a single unit of service type x. Some cost-determining factors are influenced by data center size, others by its location. For the moment, these factors are merged into and addressed as EoS (Economies of Scale) and EoL (Economies of Location). EoS and EoL express the influence of size and location on production costs. Both depend on the facility's strategy. They are zero when neither size nor location have any effect. $\text{EoL}(y) = 0.2$ means that costs of a facility with strategy y are reduced by 20% due to local effects (e.g. cheaper energy) in comparison to R_x. EoS is also increasing with facility size. Only one facility per strategy is assumed for now and a facility can only follow one strategy. $\text{EoS}(y)$ hence increases over market share of strategy y. The production costs of both service types in the facility with strategy y are defined according to Equation 5.1.

$$\begin{aligned} C_{\text{processing}}(y) &= R_{\text{processing}} \cdot (1 - \text{EoS}(y)) \cdot (1 - \text{EoL}(y)) \\ C_{\text{storage}}(y) &= R_{\text{storage}} \cdot (1 - \text{EoS}(y)) \cdot (1 - \text{EoL}(y)) \end{aligned} \tag{5.1}$$

Different service types x are reckoned as different markets (dependencies between them are explained in Section 5.4.2). The market share of a facility strategy y in the market of service type x is defined as $S_x(y)$. A market share cannot be negative. For our two markets (storage and processing) it hence holds according to Equation 5.2.

$$\begin{aligned} S_{\text{processing}}(p) + S_{\text{processing}}(c) &= 1 \quad & S_{\text{processing}}(s) &= 0 \\ S_{\text{storage}}(s) + S_{\text{storage}}(c) &= 1 \quad & S_{\text{storage}}(p) &= 0 \end{aligned} \tag{5.2}$$

Demand is modeled analog to market shares as presented in Equation 5.3.

$$\begin{aligned} D_{\text{processing}}(p) + D_{\text{processing}}(c) &= 1 \quad & D_{\text{processing}}(s) &= 0 \\ D_{\text{storage}}(s) + D_{\text{storage}}(c) &= 1 \quad & D_{\text{storage}}(p) &= 0 \end{aligned} \tag{5.3}$$

Although the demand types match the modeled strategy types, the demand of a certain type does not necessarily have to be met by a facility of the same type: Combined processing and storage demand can be met by independent p and s while a facility with strategy c might also meet independent processing and storage demand (Figure 5.1).

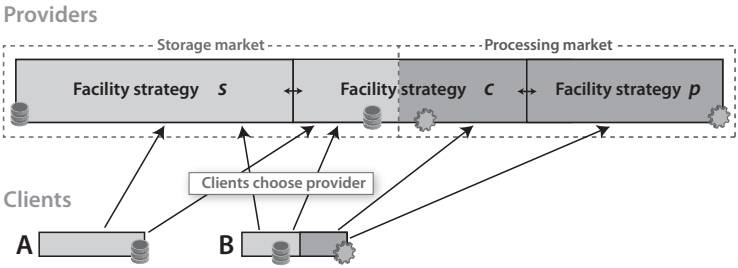


Figure 5.1: Clients are free to choose a provider for their storage and processing demand independently.

Accordingly, $D_{\text{storage}}(c)$ is the share of storage demand that is used together with $D_{\text{processing}}(c)$, regardless of where this demand is actually met. $S_{\text{storage}}(c)$, on the other hand, is the storage market share of combined facilities, no matter how it is used. The whole provisioning is not completely arbitrary, though, as facility competitiveness differs: While separate locations might feature better EoS or EoL, remote data access when combining p and s means additional transfer charges and also affects performance. We define *combined demand* for each market in order to be able to differentiate between demand that is affected by these disadvantages and demand that is not. Client B in Figure 5.1, for example, can choose to meet its storage and processing demand in different facilities. If the demand is combined demand, though, it can benefit from choosing c over s and p.

5.4.2 Fitness Functions

The fitness of each facility strategy reflects its relative commercial success in this context. Lower production costs yield more profit or allow lower prices which is more attractive for potential clients. Since the relative production cost $C_x(y)$ of service x in a facility that follows strategy y was already defined, the fitness function for s and p in the market of service x can simply be defined as 1 divided by cost (Equation 5.4).

$$\begin{aligned} F_{\text{processing}}(p) &= \frac{1}{C_{\text{processing}}(p)} & F_{\text{processing}}(s) &= 0 \\ F_{\text{storage}}(s) &= \frac{1}{C_{\text{storage}}(s)} & F_{\text{storage}}(p) &= 0 \end{aligned} \tag{5.4}$$

Unlike the strategies s and p, the fitness of strategy c is potentially raised by the savings of transfer cost or performance gains in comparison to the other strategies. This only affects demand that benefits from colocated services but

is not met by c (Equation 5.5). The constant G_x (gain) indicates the amount that a user saves by using one unit of service x in a combined center over combining separate services.

$$F_{\text{processing}}(c) = \begin{cases} \frac{1}{C_{\text{processing}}(c) - G_{\text{processing}}} & \text{when} \quad S_{\text{processing}}(c) \leq D_{\text{processing}}(c) \\ \frac{1}{C_{\text{processing}}(c)} & \text{else} \end{cases}$$

$$F_{\text{storage}}(c) = \begin{cases} \frac{1}{C_{\text{storage}}(c) - G_{\text{storage}}} & \text{when} \quad S_{\text{storage}}(c) \leq D_{\text{storage}}(c) \\ \frac{1}{C_{\text{storage}}(c)} & \text{else} \end{cases} \tag{5.5}$$

The overall gain G that a client has when it chooses a combined facility over separate service-specific facilities is split up between both markets ($G = G_{\text{processing}} + G_{\text{storage}}$). Accordingly, both market specific gains are between zero and G but cannot be appraised individually since only G as a whole is experienced. Further, because the gain only applies when a user obtains all services from c, an equal (or higher) fraction $\frac{S_x(c)}{D_x(c)}$ is required in the other market x for the first case in Equation 5.5 to apply. If the market share of c is too low in the other market, its share has to be raised in that market as well in order to gain from colocation. Client B in Figure 5.1 for example has to choose c for both storage and processing or it does not gain from colocation. Hence, the individual markets depend on each other.

5.4.3 Analysis

Following the approach of replicator dynamics [83], we consider the facility population as the player of an evolutionary game. The mixed strategy that this player pursues corresponds to the strategy distribution throughout the population (e.g. facility size), i.e. their market shares. The fitness of each facility strategy depends on the current strategy distribution. The fitness of a mixed

strategy is the weighted average of these facility strategy fitnesses. The mixed strategies in which the share of market x maximally differs by some ε of the corresponding share in a strategy m are called m's *neighborhood* in that market. The mixed strategy of a market is often referred to as a *market situation* in the following.

A mixed strategy m that has a higher fitness than any other mixed strategy n has under m's market shares is an *evolutionarily stable strategy (ESS)* [96]. A mixed strategy m is *dynamically stable*[1], when there exists a neighborhood for m so that all strategies n in m's neighborhood feature a lower fitness than m under n's market shares [83]. An ESS is also dynamically stable.

The number of market situations that can be dynamically stable is limited. When there are economies of scale, the fitness of each facility strategy increases with its market share. This means that the facility strategy with the highest fitness still has the highest fitness when its market share increases. Accordingly, a mixed strategy n that features a higher share of this strategy than a mixed strategy m also has a higher fitness at n's market shares. In consequence, no mixed strategy can be dynamically stable and only the pure strategies where one of the facility strategies has a market share of 1 remain. The only exception can occur when the share of the combined facility c is identical to its demand. Although c can have a higher fitness than s or p at a lower market share, the lost colocation gain when c's market share exceeds combined demand reduces c's fitness significantly. This means that the above argument does not apply here and this mixed strategy can be dynamically stable. Accordingly, there are up to two ESSs and potentially one other dynamically stable strategy that is not an ESS for each market. The following list presents these three potentially stable states.

[1]Dynamically stable states are often called *evolutionarily stable states* (*state* not *strategy*, e.g. [97]), but *dynamically stable* is easier to distinguish from ESS and hence preferred here.

ESS 1 All demand is met by the colocated data center.
$S_{\text{processing}}(c) = 1$ or $S_{\text{storage}}(c) = 1$

ESS 2 All demand is met by the locally separate facility.
$S_{\text{processing}}(p) = 1$ or $S_{\text{storage}}(s) = 1$

DSS Combined demand is met by the colocated facility and independent demand is met by the locally separate facility.
$S_{\text{processing}}(c) = D_{\text{processing}}(c)$ or $S_{\text{storage}}(c) = D_{\text{storage}}(c)$

Which dynamically stable strategies actually exist depends on the magnitudes of scale/location economies and colocation gain. A mixed strategy's fitness improves with a higher share of a strategy that has a better fitness. It hence is sufficient to compare the fitnesses of pure strategies in order to determine whether there is a mixed strategy that features a higher fitness than the current mixed market strategy at the present market shares.

ESS 1 exists when in a situation where c serves the entire market x ($S_x(c) = 1$), the fitness of c is higher than the fitness of s or p (depending on the market). Since a higher market share increases the fitness of the colocated facility, this is the case when the fitness of c is the highest for some market share of c that exceeds colocation demand $S_x(c) > D_x(c)$. Inequality 5.6 shows the according conditions for an ESS 1 in the processing and the storage market.

$$\begin{aligned} F_{\text{processing}}(c) > F_{\text{processing}}(p) &\Leftrightarrow C_{\text{processing}}(c) < C_{\text{processing}}(p) \\ F_{\text{storage}}(c) > F_{\text{storage}}(s) &\Leftrightarrow C_{\text{storage}}(c) < C_{\text{storage}}(s) \end{aligned} \tag{5.6}$$

ESS 2 exists when the fitness of c is the lower than the fitness of s or p in a situation where c has no market share ($S_x(c) = 0$). With the same argument as above, this is the case when it holds for some positive market share $S_x(c) < D_x(c)$. The conditions for both markets are presented in Inequality 5.7.

$$\begin{aligned} F_{\text{processing}}(c) < F_{\text{processing}}(p) \Leftrightarrow G_{\text{processing}} < (C_{\text{processing}}(c) - C_{\text{processing}}(p)) \\ F_{\text{storage}}(c) < F_{\text{storage}}(s) \Leftrightarrow G_{\text{storage}} < (C_{\text{storage}}(c) - C_{\text{storage}}(s)) \end{aligned} \tag{5.7}$$

As stated in Section 5.4.1, the market-specific colocation gain G_x cannot be evaluated individually. Because of the interdependence of the markets, a more general condition for ESS 2 has to be formulated (Inequality 5.8).

$$G < C_{\text{processing}}(c) - C_{\text{processing}}(p) + C_{\text{storage}}(c) - C_{\text{storage}}(s) \tag{5.8}$$

DSS exists is market x when the according condition (depending on the market) in Inequality 5.9 holds for some situation where c has a market share $S_x(c) > D_x(c)$ and when additionally Condition 5.10 holds for some market share $S_x(c) < D_x(c)$. Note that while these conditions are the opposites of the ESS conditions, the corresponding ESS and DSS conditions are not mutually exclusive since both might hold in different market situations (different $S_x c$). Further, note that Condition 5.10 again depends on the other market.

$$\begin{aligned} F_{\text{processing}}(c) < F_{\text{processing}}(p) \Leftrightarrow G_{\text{processing}} < (C_{\text{processing}}(c) - C_{\text{processing}}(p)) \\ F_{\text{storage}}(c) < F_{\text{storage}}(s) \Leftrightarrow G_{\text{storage}} < (C_{\text{storage}}(c) - C_{\text{storage}}(s)) \end{aligned} \tag{5.9}$$

$$G > C_{\text{processing}}(c) - C_{\text{processing}}(p) + C_{\text{storage}}(c) - C_{\text{storage}}(s) \tag{5.10}$$

To illustrate the dynamical stability of DSS, suppose the storage market in a situation where c has a lower market share than there is colocated storage demand ($S_{\text{storage}}(c) < D_{\text{storage}}(c)$) and additionally c has a higher fitness

than s. The market share of c increases due to its higher fitness and its fitness remains higher than s's fitness up to a share that equals combined demand ($S_{\text{storage}}(c) = D_{\text{storage}}(c)$). The higher fitness of c violates the conditions for an ESS (since it leads to a further increase of $S_{\text{storage}}(c)$ above $D_{\text{storage}}(c)$). In a market situation where c's share exceeds combined demand ($S_{\text{storage}}(c) > D_{\text{storage}}(c)$), on the other hand, the reduced fitness of c may cause it to be lower than s's fitness and in that case decreases c's market share until it equals combined demand. $S_x(c) = D_x(c)$ is then dynamically stable. This is further explained in Section 5.4.4. A dynamically stable strategy m has a neighborhood of strategies in which the market shares converge to m.

Since EoS(c) depends on $S_x(c)$ in other markets, the whole IaaS market is only stable when all individual markets are in a stable state. Next to the situation where both markets are in ESS 1, ESS 2 or DSS at the same time, the overall IaaS market can also be in a state where the storage respectively processing market is in ESS 1 and the other one is in ESS 2. A market can only be in DSS when $S_x(c) \geq D_x(c)$ is true for all markets (Section 5.4.1). Thus, ESS 1 and DSS might coexist in different markets, while ESS 2 and DSS cannot.

5.4.4 Development over Time

A modification of strategy shares does not necessarily require rational choice. In a growing market, a facility with a more successful strategy features faster growth than its competitors and thus also a growing market share. Although the mixed strategy of the population changes, this cannot be considered an intentional move: success is not a matter of choice. Such dynamics can be simulated by consistently changing strategy shares based on their relative fitness. Doing so, different initial market shares can lead to different stable states. The market in Figure 5.2 for example converges to DSS when a separate stor-

age facility meets a relatively low share of storage demand (left). It converges to ESS 2 when the separate storage facility has a higher initial market share (right).

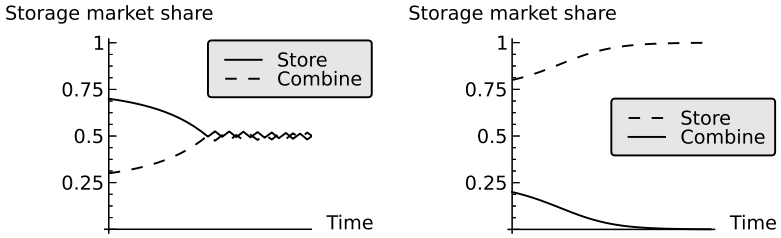


Figure 5.2: Different initial market shares result in different stable states.

EoS grows with a facility's market share, which again raises the facility's fitness. A strategy with initially better fitness enters a positive feedback loop that ultimately ends in either ESS 1 or ESS 2 in most cases. A higher fitness of strategy s respectively p results in the exclusive use of these separated facilities and a higher fitness of strategy c results in a market where c meets all demand.

There might be the case, though, that initially better fitness of the separated facility (s or p, depending on the market) reduces with growing market share despite this feedback loop. This happens when the separated facility has lower production costs than c but users that demand combined services have a higher gain from switching to c than there is a fitness difference that is caused by production costs. The fitness of c is raised and outperforms competition as soon as $S_x(c)$ drops below $D_x(c)$. As the fitness of c shrinks again when its share outgrows combined demand, the market is stuck in DSS or oscillates around it.

As costs depend on EoS and thus on market share, the cost advantage of s or p might exceed the colocation gain at very low market shares of c. The market converges to ESS 2 despite the existence of DSS in that case.

The mixed strategies at which the facility strategies (to combine or to specialize on one service) have the same fitness create thresholds between market shares that result in different stable states (Equations 5.11 and 5.12).

$$\begin{aligned} C_{\text{processing}}(c) = C_{\text{processing}}(p) \quad &\text{when} \quad S_{\text{processing}}(c) \leq D_{\text{processing}}(c) \\ C_{\text{storage}}(c) = C_{\text{storage}}(s) \quad &\text{when} \quad S_{\text{storage}}(c) \leq D_{\text{storage}}(c) \end{aligned} \tag{5.11}$$

$$G = C_{\text{processing}}(c) - C_{\text{processing}}(p) + C_{\text{storage}}(c) - C_{\text{storage}}(s) \tag{5.12}$$

If all three dynamically stable states exist for the market, both thresholds exist. Shares resulting in ESS 1 and DSS are separated by the threshold defined by Equation 5.11, Equation 5.12 separates shares leading to ESS 2 and DSS. If DSS does not exist, Equation 5.11 is never true and the second threshold separates shares that result in ESS 1 or 2. As the markets are linked, the thresholds in one market depend on the shares in the other markets.

All possible IaaS market shares can be represented in an 2-dimensional space (n-dimensional for n markets). Each dimension states the market share of a separated facility strategy, which leaves the rest of both markets to the colocated strategy. The stated thresholds divide the space in fragments that end up in a specific ESS over time (Figure 5.3). The threshold by Equation 5.11 is market-specific while the threshold by Equation 5.12 is the same for both markets. Note that Figure 5.3 only shows the thresholds for the storage market for simplicity; for the processing market, the dashed threshold would be vertical. Threshold market shares are in an equilibrium but not dynamically stable and thus very prone to disturbance, which makes them unlikely to exist long.

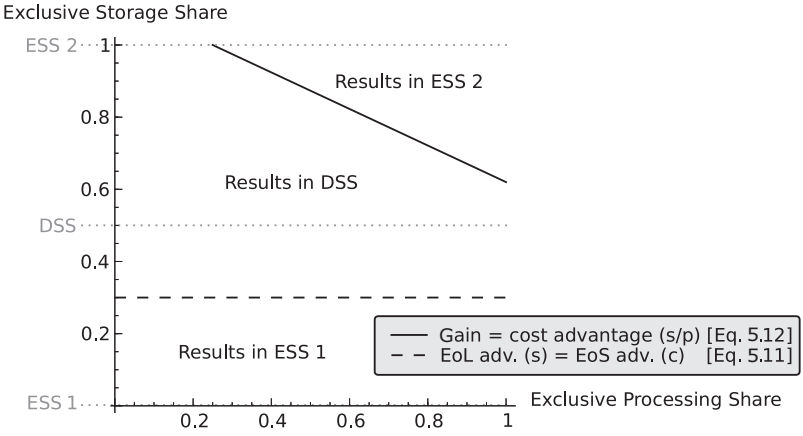


Figure 5.3: Mapping of IaaS market shares and resulting stable state in the storage market. The dotted lines indicate the market share of s after the market reaches a specific stable state.

The higher the colocation gain is compared to maximum economies of scale and location, the smaller becomes the area of shares resulting in ESS 2. The area regarding ESS 1 grows with shrinking EoL(s) as the colocation strategy needs lower EoS to compensate. When the threshold would exceed $D_s(c)$ (identical with the dotted line marking DSS) there is no DSS.

5.5 Colocation Gain Distribution

5.5.1 Gain Distribution Extension of the Model

The game model that is presented in Section 5.4 assumes a customer base that is formed by two groups: customers that gain from colocated services op-

posed to those who do not gain. While this is very likely true, the assumption that all those who benefit from a colocation of processing and storage actually have a gain of the same value is rather strong and probably too simple to adequately reflect reality. This section seeks to make the model more realistic by introducing a colocation gain distribution that can map benefits of different magnitude amongst the clients.

The colocation gain distribution amongst potential clients can be described by a *CCDF (complementary cumulative distribution function)*. Any market share of the colocated facility strategy is associated with a number of clients that have a certain colocation gain as a minimum. We define a decreasing function G_{min} that maps an IaaS market situation (with a certain market share of c for both markets: $S_{\text{processing}}(c), S_{\text{storage}}(c)$) to a specific colocation gain. All demand that is not served by the colocated facility would gain from switching to the colocated facility by the function value or more.

Although the model now regards the different client benefits of colocated services via the colocation gain distribution, this does not affect the general game setup. Each client still has to decide between the different providers and this affects the fitness of different facility types. In a situation of specific economies of scale and location to c, s and p (and accordingly different prices for the processing and storage services), the individual colocation gain of a client determines whether a colocated facility is the better option for this client. Accordingly, any situation has a threshold colocation gain that splits all potential clients into two groups: clients that are better off with a colocated facility and those who better decide for the separate facilities because their gain would be too low to compensate for the higher price of the colocated services. The threshold gain (Equation 5.13) is identical to the definition in Equation 5.12, although it separates different clients instead of market situations in this context. The inverse function of G_{min} determines the according share of overall demand that belongs to these two client groups.

$$G_{\text{threshold}} = C_{\text{processing}}(c) - C_{\text{processing}}(p) + C_{\text{storage}}(c) - C_{\text{storage}}(s) \quad (5.13)$$

For any market situation, the current market shares determine economies of scale which in turn determine the demand share of clients who have a potential benefit from using colocated services instead of separate facilities. The fitness of the different facility strategies can be calculated based on this potential by assigning the said demand share to $D_x(c)$. When the market share is smaller than the demand share of clients that have an incentive to switch to colocated facilities ($S_x(c) < D_x(c)$), the fitness of the colocated strategy is raised in the same way as defined in Equation 5.5. The gain value G_x that improves the fitness can be defined as the average gain of all clients that would benefit from colocation but are not obtaining the services by the colocated facility. Most important is the fact that the actual gain is larger than the minimum gain that is given by the distribution function ($G_x > G_{\text{min}}$).

5.5.2 Colocation Gain Dynamics and Stable States

The threshold colocation gain and the according demand share are subject to change with economies of scale and hence depend on current market shares. In consequence, the fitness of the colocated strategy changes over time and this has an impact on the stable states of the model.

The existence condition of ESS 1, where all demand is met by colocated facilities, is independent from the colocation gain and it is hence not affected by the assumption of a colocation gain distribution (Inequality 5.6).

For ESS 2, however, the colocation gain distribution has to be considered. Whenever there are no clients that have a higher gain from colocation than from the cheaper prices of the separate services, the fitness of the separate

facility strategy is higher than the colocation strategy. The threshold gain is defined as the minimum gain that causes a higher fitness of the colocation strategy in comparison to the separated strategy. Since both the threshold gain as well as the gain distribution depend on the market situation, the existence condition of ESS 2 has to be met for zero market share of c, i.e. the market situation ESS 2 itself). This implies that Condition 5.7 has to hold for the maximum colocation gain of the distribution.

The DSS case is a bit more complicated since there is not a single combined demand share as a potential stable state. An increased market share of the colocation strategy lowers its difference in EoS compared to the separate facilities. As a consequence, the threshold colocation gain is also reduced, which eventually rises the demand of the colocated service. Nevertheless, a situation where $S_x(c) = D_x(c)$ is possible (e.g. this can happen when a higher market share causes an insufficient increase of demand). No client that uses separate facilities would gain enough from colocation in order to justify a switch to colocated services. In such a situation, the function value of the colocation gain CCDF is identical to the threshold colocation gain at the current market shares. Figure 5.4 shows an example for the development of $G_{\text{threshold}}$ and G_{min} over the market share in one market (at a fixed market share in the other market). The intersection of both curves marks a situation where $S_x(c) = D_x(c)$, which might meets DSS conditions. In intervals over market share where the value of the colocation gain CCDF is higher than the threshold gain, there are clients that prefer colocation over the cheaper separated facilities. (Formally, $G_{\text{min}} > G_{\text{threshold}}$ implies $D_x(c) > S_x(c)$.) Since the fitness of the colocation strategy is improved by a value G_x that is larger than G_{min}, Condition 5.10 always holds in such a situation. In intervals where the threshold gain is larger than the CCDF, there are no clients that would gain enough from colocation ($G_{\text{min}} < G_{\text{threshold}}$ implies $D_x(c) < S_x(c)$). The fitness of the colocation strategy depends on production cost alone in such

a situation and Condition 5.9 applies. It holds where the threshold gain is positive. In consequence, there are dynamically stable states wherever G_{min} equals $G_{\text{threshold}}$ and when there is a neighborhood where $G_{\text{min}} > G_{\text{threshold}}$ holds for smaller market shares and both $G_{\text{min}} < G_{\text{threshold}}$ and Condition 5.9, hold for larger market shares. (This is true for the intersection in Figure 5.4, which hence is DSS.)

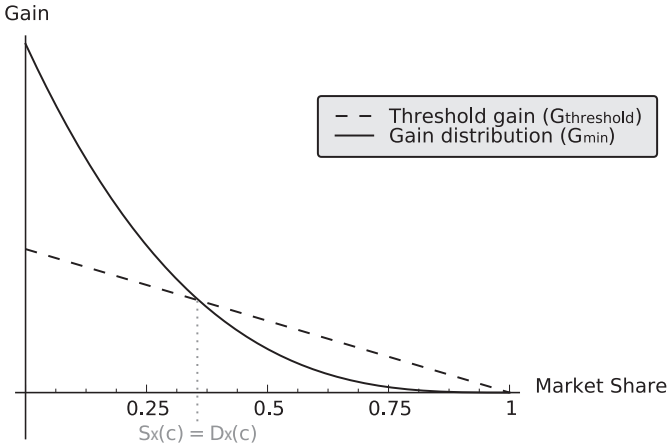


Figure 5.4: Colocation gain distribution function and threshold gain over colocation market share. The intersection marks a potential DSS.

5.6 Implications on IaaS Clouds

5.6.1 Possible Economies of Scale and Location

As discussed in Section 3.4, economies of scale of almost 20 % are realistic for a processing facility by scaling up from 1 000 servers to over 50 000. This

can be achieved by major reductions of administration effort and better power usage effectiveness. Those savings get close to optimality and only marginal further improvements can be expected. Scale economies of storage seem to be a lot better with large-scale commodity storage solutions being about six times cheaper (per GB) than storage area networks in small facilities [45]. This means possible storage EoS of over 80 %.

Potential economies of location are less complex infrastructure (e.g. cooling, uninterrupted power supply) and cheaper operating costs regarding energy consumption (infrastructure) and price in the first place. In the total cost of ownership example in [60], infrastructure cost is about 7.5 %, electricity cost about 15 % (€ 0.1 per kWh) of yearly costs of a processing facility. In a place with a free and reliable power supply and a climate that allows passive cooling (no infrastructure and energy costs), location economies of a little over 20 % would be possible. This means that the theoretic maximum of EoL is about the same as EoS. Unlike the latter, EoLs close to optimality are unrealistic. International industry energy pricing suggests that cutting costs in half is possible, so processing EoL of about 10 % might be realistic for a cool country with cheap energy. Storage EoL are negligible due to the small impact of energy and cooling on storage costs.

There also might be other economies of location that evolve from the future Internet development. For instance, the location of cloud storage facilities can be important for their use in content distribution. Such a scenario is explored in Chapter 6.

5.6.2 Stable Markets in IaaS

Section 5.4.3 presents potential stable market situations. Their existence under the EoS and EoL estimations from Section 5.6.1 is evaluated in the following.

As the data center location is important for the cost of processing but not for storage, only facilities following strategy p or c have an incentive to choose an economically interesting location. Due to legal circumstances and clients' risk awareness, an exclusive storage facility probably prefers a location that is close to the client instead. Other intangible criterions like political stability or a lower risk of natural disasters can also be considered without tradeoff against cost-reducing factors. Strategy c either chooses the location of p with high EoL for processing (Scenario 1) or the location of s in order to benefit from the intangible assets (Scenario 2). This is illustrated in Figure 5.5.

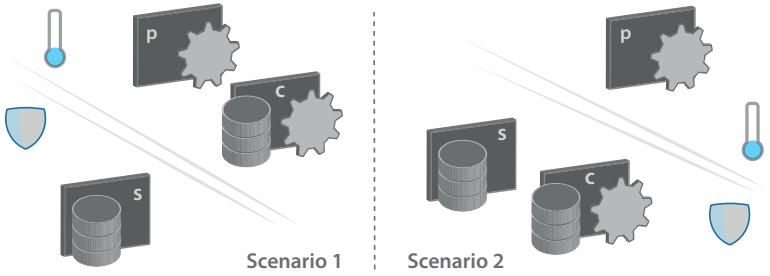


Figure 5.5: Two possible scenarios for facility placement. The colocated facility location can either optimize operation cost (e.g. by choosing a cool location) or have intangible assets like a higher security level.

Under the assumption that there are no synergetic scale economies, EoS and EoL of strategy c can be defined market independent. In Scenario 1, c has the same EoS and EoL as the separate facility (s or p, depending on the market) when both have the same market share. When c has the higher share, Inequality 5.6 is true and hence ESS 1 exists. In Scenario 2, there are no EoL of c in the processing market. ESS 1 exists nevertheless, because the possible EoS-difference of 20 % is larger than $\mathrm{EoL(p)}$ of 10 %.

An existence of ESS 2 requires the highest colocation gain to be smaller than all possible savings by EoS and EoL of the separate services (Section 5.5.2). This means that in Figure 5.4, the gain distribution function has to be smaller than the threshold gain at zero market share. The savings due to scale and location can be quite significant at very low shares of c with up to 30 % for processing and 80 % for storage in both scenarios. Clients mostly benefit from colocated services because they have better performance and there are no traffic charges. Data rates between Amazon S3 and EC2 within the same region are about 10 MB/s [49], whereas moving data from one S3 region to another is reported to be a mere 1 MB/s. Although this is more of an example than a proper evaluation and not all applications need a lot of bandwidth, it shows how massive the colocation gain can be. Latencies can also be expected to be a lot higher over some distance than in a facility's local network. Thus, ESS 2 is a possible outcome only for very small shares of c, but even its existence is quite unlikely.

In contrast, the DSS conditions are much more likely met since. It is not very probable that the savings from separate services are always higher than the colocation gain of their users. On the other hand, there are applications that do not require different services to work together (e.g. an archive) which means no or very little colocation gain for the corresponding demand share. We can therefore assume that a state exists where $S_x(c) = D_x(c)$ and where Condition 5.10 holds (Section 5.5.2). When additionally EoS and EoL of c are lower than those of s or p (Condition 5.9), the market situation is dynamically stable. Like in the case of ESS 1, the strategy with the larger share features lower costs in Scenario 1, thus DSS can only exist with a share of $S_x(c) < 0.5$. In Scenario 2, the worse location economies of c make the existence of DSS a lot more likely for processing: It exists when $S_p(c) < 0.75$ (assuming linear growing scale economies).

It is important to notice that DSS and ESS 2 are not mutually exclusive since the threshold gain can exceed the distribution function at zero market share but fall below it at higher shares. Also, depending on the gain distribution in the market and how economies of scale develop over market share, several DSS are theoretically possible.

If DSS exists, the market reaches it at initial shares of $S_x(c) < 0.5$ (respectively $S_{\text{processing}}(c) < 0.75$ in Scenario 2). If the market shares are higher or DSS does not exist, the market reaches ESS 1. At very low shares of c, a potentially existing ESS 2 could also be reached.

5.6.3 Conclusions

A market where all demand is met by colocated facilities (ESS 1) is in a stable constellation and there are no circumstances that challenge this stability.

For DSS, a demand share of $D_x(c) < 0.5$ with zero or low colocation gain might be realistic for storage, where lots of data just sits around, but processing of more than half of the available quantities without much data I/O can hardly be expected. When combined facilities consider risk-aware customers or those with legal restrictions in their site selection (Scenario 2), DSS existence requires a lower share of processing without much data access. Nevertheless, also this lower share – a quarter of the demand – is not very likely. This is reflected in the poor range of pure processing services today. In conclusion, the coexistence of storage centers and combined facilities (DSS) is a possibility while the persistence of exclusive processing centers is unrealistic (but could become an option in a very large market, see Section 5.7). Separate storage services exist in today's market with object storage like Amazon S3, which is reported to store over a trillion objects [8]. It is difficult to estimate the amount of actual storage demand, but assuming an average size of 100 kilobytes per object, this sums up to 100 petabytes. Each of the suspected

450 000 blade servers in use for EC2 [69] would require an average of 240 GB of disk space to generate the same amount of combined storage demand. This means that separate storage demand appears to be high enough in order that corresponding storage facilities are large enough to be competitive in separate locations. It depends on the amount of separate storage which actually takes place in separate facilities today, whether the market converges to a situation where these separate facilities (still) exist.

With respect to the large shares of combined services like Amazon EC2 in the current market, the possibility of a market where processing and storage takes place in completely separate facilities (ESS 2) is a rather academic option. It also requires massive improvements of latencies and bandwidth for data access over the Internet for such a stable market situation to exist.

5.7 Discussion of the Model

This section discusses further aspects of the presented model for clarification and starting points for future research.

5.7.1 Preference of Combined Demand

The fitness function of colocated facilities suggests that any demand such a facility provides is preferably combined demand. In theory, it could provide clients with independent storage and processing demands while some combined demand is still met in separate facilities. Limiting the influence of a colocation gain to $D_x(c) > S_x(c)$ underestimates the fitness of strategy c in such a case. But separated facilities feature better EoL and can offer lower charges whenever the colocated strategy does not feature better EoS of the same magnitude. As clients with independent demand do not benefit from a colocation of services, they are expected to generally prefer separated facili-

ties if they can offer lower prices. If the EoS advantage of colocated facilities is higher than the competitor's EoL advantage, this results in higher fitness of c anyways.

5.7.2 Several Facilities per Strategy

As described in Section 5.4, the mixed strategy of the player reflects market shares of the pure facility strategies. Those shares can be formed either by providers that exclusively follow one pure strategy (as modeled previously) or by providers that follow a mixed strategy. For instance, there might be one provider that operates both facilities s and p and another provider that operates facility c. This hardly affects the model presented so far. Another option, though, is the existence of several facilities of the same type that provide the share of a strategy together. In comparison to a single huge facility, this results in smaller EoS for each facility, which means a lower average fitness of this strategy. This affects the constraints that lead to specific stable state and especially reduces the likelihood that highly segmented strategies are successful. The model currently does not include unbalanced scattering of the strategies' market shares. Such scattering would affect the gradient of EoS over market share and thus alter the thresholds in Section 5.4.4. Possible EoS (Section 5.6.1) might not be reached when many facilities follow the same strategy as the market is of limited size. This could also affect the existence of the stable states.

5.7.3 Very Large Facilites

Economies of scale appear to reach a maximum at today's facility sizes (Section 3.4). In an even larger market, this results in an initial strong increase of EoS that more and more flats out over facility size (market share): The larger

the market gets, the less important do scale economies become compared to locational gains. This means for the processing market that DSS exists for even higher $D_p(c)$ and is reached at accordingly low shares of p. Assuming the initial market entry barrier of reaching this share can be taken, locally separate processing becomes more likely in the future.

5.7.4 Applications Other than IaaS

While the proposed market model is applied to IaaS in this chapter, it approaches specialized vs. diversified product strategies in general. Its adaptation to similar problems, which also may involve more than two service types, should be possible without much difficulty.

"Prediction is very difficult,
especially about the future."

– Niels Bohr

6 Cloud Infrastructure and the Future Internet

6.1 Introduction

Internet traffic has increased over the last years not only because of a growing user base, but also because data-intensive services (e.g. video streaming) have become more common [24]. The delivery of such content causes a lot of data transfer in the Internet backbone. This traffic can be reduced by caching technologies, which prevent repeated transport of the same data over long distances and also provide a better user experience due to lower latencies. While caching is of benefit to both the *Internet Service Provider (ISP)* and the *Content Provider (CP)*, they utilize caching independently. ISPs may cache for internal optimization and many CPs make use of specialized *Content Distribution Networks (CDNs)* to improve user experience.

This chapter analyzes a business model for a cooperative caching solution where ISP and CPs share the caching cost and in consequence the ISP caches more data. This reduces traffic and hence cost of the ISP's network and increases the distribution of content at the same time. The question is whether or not such a business model can prevail in the market and under what conditions it may exist.

Our studies are of high importance in a cloud context. While today cloud computing relies on the power of big data centers, several studies have already pointed out the benefits of having distributed clouds, e.g. [98]. In such environments, ISP caching and CDNs can be flexibly managed according to demand by making use of cloud storage facilities.

We propose game-theoretic models for both the feasibility of the business models as well as the potential resource allocation in the cloud. Further, we investigate the long-term incentives of such a paid ISP caching service in a third, repeated game.

The remainder of this chapter is organized as follows. Similar research is presented in Section 6.2. Section 6.3 gives important background information about caching today and describes the considered business models. Two game-theoretic models are set up in Section 6.4 in order to analyze the feasibility of the business models as well as the resource allocation in the cloud. Section 6.5 investigates Nash equilibrium and Pareto optimality conditions in these two games. The long-term incentive of CPs to pay for ISP caching is discussed in Section 6.6 and Section 6.7 concludes the chapter.

6.2 Related Work

The importance of the ISPs' involvement in cache deployment is becoming evident as CPs like Netflix, who have been using CDNs for a long time, are now deploying their own caches within the ISPs' networks [75]. Some state-of-the-art research studies the ISPs' involvement in the caching process; most however neglect the business aspects and target the study more from a resource perspective. For instance, a technical solution for ISP-driven caching that takes the role of a CDN is presented in [22], for instance. Although the system implies a business model similar to the one that we investigate in this

chapter, only the efficient use of network resources is evaluated. A similar approach is taken in [53], where the placement, size and number of caches is analyzed in an ISP-operated CDN.

Several papers study possible cooperations between ISPs and CPs at a control level. The authors in [52] look into different approaches that an ISP can take in managing traffic engineering and server selection, ranging from running the two systems independently to designing a joint system. The surprising conclusion from this work is that in the case of two independent systems, extra visibility between the two systems results in a less efficient outcome. Server selection and traffic engineering is also studied in [30]. Although these publications study the cooperation between ISPs and CPs, it is important to highlight that their focus is on cooperation from a control perspective, while ours is on a cooperation in cost of the caching system.

There is also important research that investigates other interactions in the market. Reference [32] studies how the cooperation between ISPs can influence transit traffic costs with respect to the cached content in a scenario where ISPs have caching capabilities. Two game-theoretic models for cooperative caching are put forward: one where the ISPs follow a selfish strategy and another where the interests of the neighboring ISPs are also taken into account. The results show that by cooperating, ISPs can achieve considerable gains, even if they follow a selfish strategy. The gains can further increase when also taking the neighboring ISPs' interests into consideration. The authors in [56] also aim at a more efficient routing by the combination of a non-uniform bandwidth pricing by the ISP with a CDN-side cost-aware routing. Other work focuses on the self-interaction within ISPs or CPs themselves, e.g. [94], [66] and [95].

6.3 Background

The main benefit of caching lies in the reduction of redundant traffic, which is caused by repeated requests and delivery of the same data. Especially for the ISPs, caching means less transit or peering costs due to reduced traffic volumes flowing outside their networks. In this chapter, traffic flowing outside an ISP's network is defined as distant traffic, whereas traffic within the ISP's network is considered as local traffic. In addition, caching may also reduce local traffic, however, due to its minor significance compared to distant traffic costs, this chapter ignores this effect.

From a CP's perspective, caching may reduce latency for its end users due to the proximity of the cache servers to the end users. Similarly, the end user perceives the ISP's service quality to improve with caching.

As a consequence of the increasing importance of caching, different existing caching technologies are operating in parallel. In addition, new technologies that utilize caching are being developed. This section gives a brief introduction to the caching technologies and explains the assumptions adopted in this chapter.

6.3.1 Caching Technologies

Web Caching

Web caching can be considered as the first caching technology in the market. The demand for web caching became evident over a decade ago [13] when the usage of the World Wide Web increased dramatically. The idea is to temporarily cache web sites in the proxy servers or in end users' browsers to more efficiently serve the subsequent requests. In addition to the ISP controlled proxy caching, web caches can also be deployed by the content providers closer to the content servers to offer origin server load balancing. The difference com-

pared to CDNs and cloud storage lies in the web caching being limited to web pages, as well as the temporary nature of the cache storage. Furthermore, web caching is done in a transparent manner, which means that the ISPs cache web pages without any agreements with the CPs.

Content Delivery Networks

A CDN [29] operates as an overlay to the basic Internet and divides the end-to-end connection into two pieces: one between the CP and the CDN servers, the other between the CDN servers and the end users. The CDN provider co-locates its data centers into the ISP's network and caches the CP's content based on the contract type: either the content is cached after it is requested for the first time or the CP can push certain content into the cache before it is requested [29].

Due to the traditionally high prices of CDNs, CPs typically use CDNs to serve only the heavy or time-sensitive content, whereas other content is served by the ISP from the origin server. However, the situation is changing as the CDN prices are dropping [86]. In addition, the proportion of data that is not served by the CDNs is negligible compared to the heavy content that is delivered through the CDNs. Thus, this chapter assumes that when CDNs are used, all content is served from the CDN servers.

Traditionally, CDNs [101] are operated by third-party CDN providers, which are here called pure-play CDNs. In addition, CDNs used to have settlement-free peering agreements with smaller ISPs for co-locating the data centers [34]. However, the relationships are changing and ISPs are increasingly charging CDN providers for the co-location service [67][103]. Other changes are also taking place today: for example, ISPs and CPs are increasingly building their own CDN networks [99][11][43][36]. As a response, the pure-play CDN providers are offering CDN licenses to ISPs [85]. In addition, the

CDN providers are working towards interconnectivity between themselves through initiatives such as CDNi [84].

Cloud Storage

Cloud computing [102] is a paradigm for better and easier hardware and software management. Clouds are pools of virtualized resources, such as software, hardware and services, that can be easily accessed. The idea of the cloud is to move the infrastructure to the network, which reduces the costs of resource management and offers better scalability and flexibility.

The cloud paradigm offers mainly three service categories: IaaS, platform as a service (PaaS) and software as a service (SaaS). In the caching case, only IaaS is relevant, where a *Cloud Storage Provider (CSP)* virtualizes its resources so that they can be split and assigned dynamically to the customers. Typically, the customer is charged only for the actually used storage and the service level agreements (SLAs) guarantee the quality of service.

In-network Caching

Furthermore, in-network caching schemes, such as ISP-driven caching and information-centric networking [3], are widely researched. With in-network caching, the content is cached in the network elements, e.g. routers and servers, when it passes through.

In ISP-driven caching, the ISPs place cache servers or caching enabled routers into their own network and cache the content either transparently or according to agreements with CPs. In addition, the ISP can choose to utilize a third-party storage provider (e.g. a CSP) or build their own caching infrastructure. Information-centric networking operates with a similar concept. However, in information-centric networking, routing is done based on content names instead of host addresses [3].

6.3.2 Assumptions

This chapter assumes a simplified content delivery ecosystem with only CPs, ISPs, CDNs, CSPs and end users, the value network of which is illustrated in Figure 6.1. The value network shows the exchanges between each of the stakeholders divided into 1) content transfer, 2) monetary transfer and 3) intangible benefits. The two caching schemes considered in this chapter are ISP-driven caching and pure-play CDNs due to their high impact in the current content delivery market.

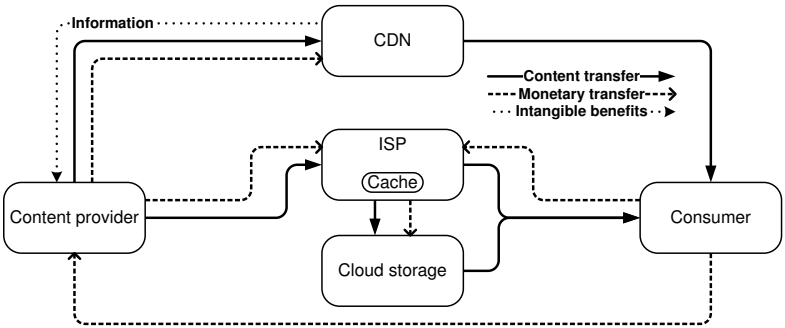


Figure 6.1: Value network.

From an ISP's perspective, it has two strategic decisions regarding caching. First, it has to decide whether to cache or not. If the ISP decides to cache, it has to decide how to price the caching service. The third decision relates to whether to buy caching services from a third party or build their own caching infrastructure. The first two decisions combine into three business models: 1) Basic service, 2) ISP internal network optimization and 3) ISP-driven caching service. The three business models are briefly explained here.

Business Model 1 – Basic service: The first business model represents a situation in which the ISP decides not to deploy caching and stays in its traditional market: access provision and traffic transmission. In this business model, the ISP charges the CPs only for the network access and offers a best-effort service. If the CP wishes to improve the *Quality of Experience (QoE)* to its end users, it can either deploy its own caching system or buy the service from a CDN. For example, Google is a content provider that has its own caching system [43] and MTV Networks uses Akamai's services [4].

Business Model 2 – ISP-internal network optimization: Business Model 1 does not fully comply with the current situation of the network, because most ISPs employ caching at some level, e.g. web caching. Thus, the second business model explains a situation in which the ISP caches content but does not charge CPs for the caching service. The incentive for the ISP is in reducing costs through optimizing its own network and reducing transit traffic volume. In this situation, the CP pays the ISP for the network access and the traffic volume in the traditional way. The difference compared to Business Model 1 is that the CP does not have a direct relationship with the CDNs, though the actual caching could also be outsourced to CDNs or CSPs.

Business Model 3 – ISP-driven caching service: In the third business model, ISPs are offering caching services to the CPs for an extra fee. The CPs can be charged based on the amount of cached data, the traffic volume generated by the caches or the combination of the two. In all three charging models, the CP contracts only with the ISP. However, the actual caching could be done by a third party as explained above. We assume that the ISP charges by bandwidth since the ISP pays for the

bandwidth required for transit as well and ISPs are trying to save in transit costs by caching the content.

We assume the existence of peering or transit agreements between multiple ISPs, which allow one ISP to offer caching services that go beyond its network. Thus, the CP has a business relation with only one ISP, which we can assume to be its local one. In addition, this chapter assumes that the ISPs do not have existing caching infrastructure and need to build the caching service before offering it to the CPs.

6.4 Game-Theoretical Setup

This section provides two separate game models that represent the interaction of ISPs and CPs regarding the different business models and the resource allocation interaction between ISPs and CSPs.

6.4.1 Business Model Game

In the following, we want to identify conditions under which the business models that are presented in Section 6.3 can exist as an equilibrium in a market situation.

A simple two-player-game with ISP and CP as the players is set up to compare the different payment models. The ISP can choose to either route all data requests to the CP or to install caches and meet requests from there. (We neglect the time that is needed to deploy the caching service for now and address this issue in Section 6.6.) The CP can choose between a traditional Internet service (Business Models 1 and 2), a service that involves a payment for caching (Business Model 3) and the utilization of a CDN (ISP's competition for caching). The resulting situation depends on the decisions of both

parties. Each decision combination features a specific utility for each player as a measure of how valuable the resulting situation is (higher is better).

Figure 6.2 depicts the normal form of the game setup. The upper left corner represents Business Model 1, the upper right represents Business Model 2. In the middle left, the CP is willing to pay for caching but the ISP does not decide to cache. Business Model 3 can be found in the middle right. U and V denote CP's and ISP's utilities.

		ISP: Don't Cache	ISP: Cache
CP	Traditional	V_1 / U_1	V_2 / U_2
	Pay for Caching	V_3 / U_3	V_4 / U_4
	CDN	V_5 / U_5	V_6 / U_6

Figure 6.2: Utility matrix of the payment model game. Each decision combination results in a situation of specific value to CP and ISP.

All utilities depend on billing: the ISP prefers a higher payment by the CP while for the CP a lower payment is more valuable. The ISP's utility is reduced by any operational expenses. In the following, the utility functions for the different outcomes of the game model are presented.

The traditional Internet service payment depends on the required bandwidth b. We assume this bandwidth to consist of a short routing distance component b_{local} and a long routing distance component b_{distant} ($b = b_{\text{local}} + b_{\text{distant}}$). Both b_{local} and b_{distant} cause costs in the local network (c_{local}), while b_{distant} also causes additional transit costs (c_{distant}).

When the ISP is caching, all distant demand can be met without transit traffic by a nearby cache. In reality, some amount of data is most likely delivered over the full distance until it is cached. We consider this amount of traffic as negligible compared to the overall traffic. Hence, in the case of caching, the requests that are otherwise associated with b_{distant} never reach the CP but are entirely served from a local cache.

The ISP charges p_{isp} as a price for bandwidth. In Business Model 1, this utility is decreased by costs in the local network and long-distance transfer (Equation 6.1). In Business Model 2, the ISP has to pay for storage instead of long-distance transfer, c_{storage} is the cost of data hosting (Equation 6.2). The CP only pays for b_{local} and benefits from better QoE. QoE^+ is the value of this improvement to the CP.

$$\begin{aligned} U_1 &= -b \cdot p_{\text{isp}} \\ V_1 &= b \cdot (p_{\text{isp}} - c_{\text{local}}) - b_{\text{distant}} \cdot c_{\text{distant}} \end{aligned} \tag{6.1}$$

$$\begin{aligned} U_2 &= -b_{\text{local}} \cdot p_{\text{isp}} + \text{QoE}^+ \\ V_2 &= b_{\text{local}} \cdot p_{\text{isp}} - b \cdot c_{\text{local}} - c_{\text{storage}} \end{aligned} \tag{6.2}$$

The CDN option comes with a service fee for bandwidth p_{cdn} and storage $p_{\text{cdn}-\text{storage}}$. We assume that all data requests (including local requests) are met by the CDN and the CP does not obtain any bandwidth from the ISP directly. Similar to ISP caching, all demand can be met from a local CDN server without the long-distance transfer. This provides a good user experience irrespective of the ISP's action. We further assume the QoE to be the same as with ISP caching. The ISP saves long-distance transfer costs and we assume that it charges the same bandwidth price from the CDN as it usually charges from the CP (Equation 6.3). Caching when the content is already distributed over the network obviously only causes costs to the ISP without any benefit

for either parties (Equation 6.4).

$$\begin{aligned} U_5 &= -b \cdot p_{\text{cdn}} - p_{\text{cdn-storage}} + \text{QoE}^+ \\ V_5 &= b \cdot (p_{\text{isp}} - c_{\text{local}}) \end{aligned} \tag{6.3}$$

$$\begin{aligned} U_6 &= U_5 \\ V_6 &= V_5 - c_{\text{storage}} \end{aligned} \tag{6.4}$$

When the service stipulates a caching payment, this does not imply that the ISP actually decides to cache. As the payment is usage-based, no caching fees have to be paid when the ISP is not caching. In this case, the utilities for the service are the same as those of the traditional service when we assume the same bandwidth price (Equations 6.5).

In Business Model 3, where caching actually takes place, utilities are based on the service fee for cache bandwidth p_{caching} that the ISP charges the CP.

$$\begin{aligned} U_3 &= U_1 \\ V_3 &= V_1 \end{aligned} \tag{6.5}$$

$$\begin{aligned} U_4 &= U_2 - b_{\text{distant}} \cdot p_{\text{caching}} \\ V_4 &= V_2 + b_{\text{distant}} \cdot p_{\text{caching}} \end{aligned} \tag{6.6}$$

6.4.2 Resource Allocation Game

When an ISP adopts caching, it can either build its own caching infrastructure or buy the caching capability from third parties. These third parties include traditional hosting service providers and CSPs. This section investigates using game theory whether the utilization of third-party hosting or the operation of own caching facilities is more feasible.

The basic idea is that third parties can offer storage cheaper due to better economies of scale, especially compared with smaller ISPs, but the ISPs may have to compromise on the cache location. Storage providers, on the other hand, might consider this possibility in their data center site selection in order to gain ISPs as their customers. However, in this chapter, despite the reduced flexibility in location choices, the caching system from a third party is assumed to offer the same QoE for end users as an ISP's own caching system. (Section 6.5.6 discusses the impact on the analysis results when this assumption is not made.) In addition, if the third party is cheap enough and savings over an own caching facility exceed extra traffic costs, the ISP has an incentive to use the third-party storage provider. Both the ISP and the third-party provider have to benefit from a situation, where the third party is involved in ISP caching or this is not likely to happen.

Another two-player-game with ISP and a third-party CSP as the players is set up. As discussed in Section 6.3, the ISP might operate own storage equipment or utilize third-party facilities (e.g. a cloud storage service). The CSP can either optimize economies of scale or partition its facilities in order to place them in several locations within the ISP's network that are more appropriate for caching. Figure 6.3 shows the normal form of the game. W denotes the CSP's utility.

The utilities depend on the price that is asked for some amount of storage, storage costs and the amount of stored data. The model distinguishes the data that the CSP stores for other clients d_{tp} and the amount of data that the ISP caches d_{isp}. Storage in its own facilities costs the ISP an amount of c_{storage}. The CSP asks $p_{\text{tp-storage}}$ for storage and the production cost of CSP storage is denoted by $c_{\text{tp-storage}}$. When the location is chosen for best size, the production costs of the CSP are reduced due to economies of scale. The cost reduction compared to the location that is best for caching is represented by the coefficient EoS. For instance, $\text{EoS} = 0.9$ means a 10 percent decrease of production

		ISP	
		Own Infrastructure	Third Party
CSP	Best Size	V_7 / W_7	V_8 / W_8
	Best Location	V_9 / W_9	V_{10} / W_{10}

Figure 6.3: Utility matrix of the cache hosting game.

cost, EoS $= 1$ means that no economies of scale apply. Accordingly, the CSP's utility varies with the location of its facilities, while the ISP's utility is not affected as long as the ISP uses its own storage facilities (Equations 6.7 and 6.8).

$$\begin{aligned} W_7 &= d_{\text{tp}} \cdot (p_{\text{tp-storage}} - c_{\text{tp-storage}} \cdot \text{EoS}) \\ V_7 &= -d_{\text{isp}} \cdot c_{\text{storage}} \end{aligned} \tag{6.7}$$

$$\begin{aligned} W_9 &= d_{\text{tp}} \cdot (p_{\text{tp-storage}} - c_{\text{tp-storage}}) \\ V_9 &= -d_{\text{isp}} \cdot c_{\text{storage}} \end{aligned} \tag{6.8}$$

When the CSP serves the ISP for caching, the additional demand increases the CSP's revenue (Equations 6.9 and 6.10).

$$\begin{aligned} W_8 &= (d_{\text{tp}} + d_{\text{isp}}) \cdot (p_{\text{tp-storage}} - c_{\text{tp-storage}} \cdot \text{EoS}) \\ V_8 &= -d_{\text{isp}} \cdot p_{\text{tp-storage}} - b_{\text{distant}} \cdot c_{\text{tp-transfer}} \end{aligned} \tag{6.9}$$

$$\begin{aligned} W_{10} &= (d_{\text{tp}} + d_{\text{isp}}) \cdot (p_{\text{tp-storage}} - c_{\text{tp-storage}}) \\ V_{10} &= -d_{\text{isp}} \cdot p_{\text{tp-storage}} \end{aligned} \tag{6.10}$$

Instead of the cost for own equipment, the ISP pays a service fee. The ISP has additional network transfer cost $c_{\text{tp-transfer}}$ when the CSP chooses to optimize economies of scale and places facilities in a relatively remote location. This affects all distant data requests, which is b_{distant}. We assume that the CSP's facility is still close enough, though, to have no significant impact on the quality of experience to the end user. We also assume that the ISP does not charge the CSP for cache-related traffic.

6.5 Game Analysis

Section 6.4 presented two game-theoretic models regarding ISP-driven caching. This section identifies the conditions for Nash equilibria and Pareto-efficient outcomes in these games.

6.5.1 Equilibria in the Business Model Game

A Nash equilibrium describes a situation where no player can unilaterally deviate to any better outcome (Section 2.2). Now, we examine which outcomes can be equilibria in the business model game and under which conditions.

(CDN, Don't Cache) is an equilibrium if $U_5 \geq U_1 = U_3$ (CP should not deviate) and if $V_5 \geq V_6$ (ISP should not deviate). The second is always true because caching facilities do have a cost, i.e. $c_{\text{storage}} \geq 0$. The first condition requires that the CP values the QoE improvement more than the additional costs incurred to the CDN compared to the traditional service (Inequality 6.11).

$$\begin{aligned} & U_5 \geq U_1 \\ \Leftrightarrow \quad & \text{QoE}^+ \geq b \cdot (p_{\text{cdn}} - p_{\text{isp}}) + p_{\text{cdn-storage}} \end{aligned} \tag{6.11}$$

Business Model 1 (traditional service without caching) is an equilibrium if $U_1 = U_3 \geq U_5$ and $V_1 \geq V_2$. The first condition is exactly the opposite as before; it holds when Inequality 6.11 is false). Second, the caching of a data object by the ISP is more expensive than the difference between transfer costs of associated bandwidth and the loss of sales of this bandwidth (Inequality 6.12).

$$\begin{aligned} & V_1 \geq V_2 \\ \Leftrightarrow \quad & c_{\text{storage}} \geq b_{\text{distant}} \cdot (c_{\text{distant}} - p_{\text{isp}}) \end{aligned} \tag{6.12}$$

Business Model 2 (caching without charging) is an equilibrium if $U_2 \geq U_4$, $U_2 \geq U_6$ and $V_2 \geq V_1$. The last condition is again the opposite of what we had before, and it is satisfied when 6.12 does not hold. The first condition is satisfied trivially since the price of caching is positive, i.e. $p_{\text{caching}} \geq 0$. The second condition requires that the ISP's price for local bandwidth is smaller than the overall price for the CDN (Inequality 6.13).

$$\begin{aligned} & U_2 \geq U_6 \\ \Leftrightarrow \quad & b \cdot p_{\text{cdn}} + p_{\text{cdn-storage}} \geq b_{\text{local}} \cdot p_{\text{isp}} \end{aligned} \tag{6.13}$$

We expect this to always hold since we assumed that the CDN provider has to pay the ISP's bandwidth price. If the CDN provider were to set the prices so low that Inequality 6.13 holds, then its service would not be profitable.

Business Model 3 is an equilibrium if $U_4 \geq U_2$, $U_4 \geq U_6$ and $V_4 \geq V_3$. The first condition never holds and thus Business Model 3 cannot be an equilibrium. However, the last condition is satisfied when the profit that the ISP might have from charging the CP directly for the bandwidth is smaller than the profit from caching. This is the case when the cost for storage is lower than

the increase in revenue (Inequality 6.14).

$$\begin{aligned} & V_3 = V_1 \leq V_4 \\ \Leftrightarrow \quad & c_{\text{storage}} \leq b_{\text{distant}} \cdot (c_{\text{distant}} - p_{\text{isp}} + p_{\text{caching}}) \end{aligned} \tag{6.14}$$

Finally, (Pay for Caching, Don't Cache) can be an equilibrium if $U_3 \geq U_5$ (Condition 6.11 does not hold) and $V_3 \geq V_4$ (Condition 6.14 does not hold). This happens when the ISP prefers not to cache over Business Model 3 and the CP prefers the ISP services over the CDN. This outcome is equivalent to Business Model 1 when considering the utilities and the caching situation, although it is maybe better regarded as a failed Business Model 3 since the CP apparently intents to benefit from caching but caching does not take place.

We can make the following observations. Whenever the ISP prefers Business Model 2 over 1 (Inequality 6.12 is not met), it also prefers Business Model 3 over (Caching, Don't Cache) (Inequality 6.14). Whenever the ISP prefers (Caching, Don't Cache) over Business Model 3 (Inequality 6.14 is not met), it prefers Business Model 1 over 2 (Inequality 6.12). Business Models 1 and 3 can be preferred by the ISP at the same time (at different CP actions), which is maybe the most interesting case (see 6.5.5).

6.5.2 Pareto Optimality in the Business Model Game

An outcome is called *Pareto-optimal* or *Pareto-efficient* when no other outcome can be found that would improve one player's utility without making the others worse off [81]. We examine under which conditions Business Model 3 is Pareto-optimal.

Pareto optimality requires that there is no (U_i, V_i) such that $U_i \geq U_4$ and $V_i \geq V_4$ and at least one of the inequalities should be strict. A sufficient condition is that $U_i < U_4$ or $V_i < V_4$ for all $i \neq 4$. Since $V_2 < V_4$ and (CDN,

Don't Cache) Pareto-dominates (CDN, Cache) (since $U_5 = U_6$ and $V_5 \geq V_6$), we do not need to consider cases when $i = 2$ or $i = 6$. We also do not need to consider case $i = 3$ since the utilities are the same as when $i = 1$. Thus, Business Model 3 is Pareto-optimal when Business Model 1 and (CDN, Don't Cache) offer a lower utility than Business Model 3 to either party: ($U_1 < U_4$ or $V_1 < V_4$) and ($U_5 < U_4$ or $V_5 < U_4$).

The condition for $V_1 < V_4$ was presented in Inequality 6.14. $U_1 < U_4$ holds when the improved user experience is more valuable to the CP than the difference of caching and Internet service price (Inequality 6.15). This depends on the individual case and some CPs may meet this condition at given prices while others do not value user experience enough.

$$\begin{aligned} & U_1 < U_4 \\ \Leftrightarrow\ & \mathrm{QoE}^{+} > b_{\text{distant}} \cdot (p_{\text{caching}} - p_{\text{isp}}) \end{aligned} \tag{6.15}$$

For $V_5 < V_4$, the difference in revenue from caching and Internet service has to exceed the caching cost (Inequality 6.16).

$$\begin{aligned} & V_5 < V_4 \\ \Leftrightarrow\quad & c_{\text{storage}} < b_{\text{distant}} \cdot (p_{\text{caching}} - p_{\text{isp}}) \end{aligned} \tag{6.16}$$

$U_5 < U_4$ holds when the charges by the ISP are lower than those of the CDN (Inequality 6.17). With the bandwidth payment model, this requires the CDN's price for storage to be at least as high as the difference between the bandwidth prices of the ISP and the CDN.

$$\begin{aligned} & U_5 < U_4 \\ \Leftrightarrow\quad & b_{\text{distant}} \cdot (p_{\text{caching}} - p_{\text{cdn}}) - b_{\text{local}} \cdot (p_{\text{cdn}} - p_{\text{isp}}) < p_{\text{cdn-storage}} \end{aligned} \tag{6.17}$$

Business Model 3 is Pareto-optimal when either Condition 6.14 or Condition 6.15 holds and additionally either Condition 6.16 or Condition 6.17 is met. This can be achieved with an appropriate pricing for ISP caching. For instance, under the assumption that the CDN charges more for bandwidth than the ISP charges for its Internet service ($p_{\text{isp}} \leq p_{\text{cdn}}$), Condition 6.17 always holds when the ISP charges no more for cache bandwidth than the CDN ($p_{\text{caching}} \leq p_{\text{cdn}}$) and at that price Condition 6.14 holds for CPs that require a distant bandwidth sufficiently high. Some CPs with a lower bandwidth may instead value user experience enough to meet Condition 6.15. Further, note that when Condition 6.16 is met, Condition 6.14 is also met.

6.5.3 Equilibria in the Resource Allocation Game

Now, we determine equilibrium outcomes and their conditions in the resource allocation game from Section 6.4.2.

(Best Size, Own Infrastructure) is in equilibrium when $W_7 \geq W_9$ and $V_7 \geq V_8$. The first conditions holds when there actually are economies of scale compared to the location that is best for caching (Inequality 6.18). The second condition is met when the additional transfer cost of cached data is higher than the savings from cheaper storage (Inequality 6.19).

$$\begin{aligned} & W_7 \geq W_9 \\ \Leftrightarrow \quad & \text{EoS} \leq 1 \end{aligned} \tag{6.18}$$

$$\begin{aligned} & V_7 \geq V_8 \\ \Leftrightarrow \quad & b_{\text{distant}} \cdot c_{\text{tp-transfer}} \geq d_{\text{isp}} \cdot (c_{\text{storage}} - p_{\text{tp-storage}}) \end{aligned} \tag{6.19}$$

An equilibrium in (Best Location, Third Party) requires $W_{10} \geq W_8$ and $V_{10} \geq V_9$. For the first condition, there must be no economies of scale to the

CSP (Inequality 6.20). Second, the price for CSP's storage has to be cheaper than own storage facilities (Inequality 6.21).

$$\begin{aligned} & W_{10} \geq W_8 \\ \Leftrightarrow \quad & \text{EoS} \geq 1 \end{aligned} \tag{6.20}$$

$$\begin{aligned} & V_{10} \geq V_9 \\ \Leftrightarrow \quad & c_{\text{storage}} \geq p_{\text{tp-storage}} \end{aligned} \tag{6.21}$$

(Best Size, Third Party) is in equilibrium when $V_8 \geq V_7$ and $W_8 \geq W_{10}$. These conditions are the opposites of Conditions 6.19 and 6.20.

(Best Location, Own Infrastructure) is in equilibrium when $V_9 \geq V_{10}$ and $W_9 \geq W_7$, which are the opposites of Conditions 6.21 and 6.18.

Under the assumption that there are economies of scale to the CSP, Condition 6.18 always holds and Condition 6.20 never holds, which means that the best size strategy dominates the best cache location strategy of the CSP. Accordingly, which of the two resulting outcomes is in equilibrium solely depends on whether the CSP service is cheap enough to outweigh the additional transfer costs (whether Condition 6.19 holds or not).

6.5.4 Pareto Optimality in the Resource Allocation Game

Section 6.5.3 presented that there cannot be an equilibrium where the CSP chooses cache-appropriate facility locations. We now investigate whether (Best Location, Third Party) can be Pareto-optimal (as introduced in Section 6.5.2).

Pareto optimality requires $W_i < W_{10}$ or $V_i < V_{10}$ for all $i \neq 10$. Unilateral changes away from (Best Location, Third Party) cause lower utilities to either the ISP or the CSP by definition: $W_8 < W_{10}$ holds since the ISP's utility

is reduced by additional network transfer cost and $V_9 < V_{10}$ holds because when the ISP does not use the CSP's service, the CSP has less revenue.

Accordingly, (Best Location, Third Party) is Pareto-optimal when $W_7 < W_{10}$ or $V_7 < V_{10}$. For the first condition, the profit from the service usage by the ISP has to be higher than the additional costs due to the sacrificed economies of scale (Inequality 6.22). The second condition is met, when the price for CSP storage is cheaper than own storage facilities (Inequality 6.23).

$$\begin{aligned} & W_7 < W_{10} \\ \Leftrightarrow \quad & d_{\text{isp}} \cdot (p_{\text{tp-storage}} - c_{\text{tp-storage}}) > d_{\text{tp}} \cdot c_{\text{tp-storage}} \cdot (\text{EoS} - 1) \end{aligned} \tag{6.22}$$

$$\begin{aligned} & V_7 < V_{10} \\ \Leftrightarrow \quad & c_{\text{storage}} > p_{\text{tp-storage}} \end{aligned} \tag{6.23}$$

Since the CSP's service price is the only common factor of these conditions, the outcome can be Pareto-optimal whenever the other factors allow a price that fulfills both conditions. Such a price exists whenever caching data in ISP facilities over CSP facilities causes extra cost that is larger than the extra cost the CSP has, when is stores all data of other customers at the lower economies of scale of a caching-friendly location (Inequality 6.24).

$$d_{\text{isp}} \cdot (c_{\text{storage}} - c_{\text{tp-storage}}) > d_{\text{tp}} \cdot c_{\text{tp-storage}} \cdot (\text{EoS} - 1) \tag{6.24}$$

This condition holds if the amount of cached data d_{isp} is sufficiently high.

6.5.5 Discussion of the Business Model Game

The analysis of the business model game shows that ISP caching where CPs participate in costs (Business Model 3) cannot be in equilibrium, which discourages the business model of selling ISP-operated caching. On the other

hand, there are possible equilibria in most other outcomes with different conditions for costs, prices and the ratio of local-to-distant bandwidth. When these conditions are met, Business Model 3 appears even less likely.

Suppose that the CP prefers Business Model 1 over the CDN. When the ISPs would offer a lower bandwidth price p_{isp} for the service that includes a cache payment, this gives CPs an incentive to switch to this service. Accordingly, Business Model 1 cannot be an equilibrium. Suppose that the ISP prefers Business Model 3 over not to cache and Business Model 1 over Business Model 2 (caching is only reasonable when CP pays). Then either the CP or the ISP has an incentive to switch strategies in the four upper market outcomes in Table 6.2 (counter-clockwise loop), which means that none of these states is in equilibrium.

Another possibility to encourage CPs to pay for caching could be the complete abolition of the traditional service. CPs would have to change their ISP in order to make a contract where they are not charged for caching. Since the CPs would thereby loose the benefits from the caching service, this would also prevent the incentive to deviate from Business Model 3. On the other hand, not offering the traditional service option could drive away many CPs. Section 6.6 investigates how Business Model 3 can be supported while holding onto both contract options.

Unstable market conditions where the ISP is not constantly caching create fluctuating QoE and cache payment. Assuming that CPs and ISPs want to avoid such discontinuities, they will likely agree on a convenient outcome. When it is Pareto-optimal, Business Model 3 can be reasonable despite not being in equilibrium since it prevents any temporary market situations at the expense of either party. Even in case of other equilibria, e.g. in favor of the CDN, the improved utility that the Pareto-optimal outcome might offer to both ISPs and CPs would encourage such a cooperation. Note, however, that maybe not all CPs meet the Pareto-optimality conditions at the same time.

6.5.6 Discussion of the Resource Allocation Game

Although (Best Location, Third Party) cannot be an equilibrium of the resource allocation game, it can be Pareto-optimal (Section 6.5.4). Therefore, the use of third-party storage for ISP-operated caching is an option that might influence the placement of these storage facilities within the ISP network.

The CSP has no incentive to switch its facility location when it expects the ISP to anyways utilize a remote CSP's service, i.e. (Best Size, Third Party). In (Best Location, Own Infrastructure), on the other hand, the ISP possibly has an incentive to switch to the CSP's service. Accordingly, the CSP is in charge to induce the desired situation. Since storage facilities cause huge capital expenditures and actually switching back and forth can be very expensive, agreements (e.g. long-term contracts) foregoing accomplished facts are advisable.

We assumed that the third party location has no effect on end-user experience. However, it possibly suffers to some extent when a remote, size-optimized CSP's facility is used for caching. A noticeable QoE decrease could negatively influence the ISP's utility in Inequality 6.9 because it weakens the ISP's selling point of a service with caching. This changes the equilibrium conditions and (Best Size, Own Infrastructure) becomes increasingly preferable over (Best Size, Third Party) for the ISP the worse QoE become. In contrast, the Pareto optimality of the cache-friendly location strategy is not negatively influenced by such a utility change.

The investment in storage facilities is an especially important factor regarding the assumption that ISPs build the caching service on-demand. In case of an own infrastructure, the deployment and modifications of e.g. capacity take some time during which the CPs would have to wait for the service to be operational. This is further discussed in Section 6.6.

6.6 After Cache Deployment

6.6.1 Storage Elasticity and Problem Description

The resource allocation game introduced in Section 6.4 investigates whether an ISP should use third-party storage or build its own facilities for caching. These two possibilities not only differ in price and location but also in flexibility. This has some important consequences once the caches are installed, which are investigated in this section.

When the ISP uses elastic cloud storage, the amount of cache storage can be scaled according to the demand. Most important, caches can be abolished once a CP stops paying for them. (The third party can then use the free capacity for other clients.) When the ISP invests in its own caching facilities, though, the ISP has an incentive to use all the installed capacity irrespective the payment by the CP since they reduce the network cost. This becomes apparent when we change the the business model game in Section 6.4 to cover a setting where the ISP has already set up its own caching infrastructure. In that case, c_{storage} also applies in V_1, which causes $V_2 > V_1$ when $c_{\text{distand}} > p_{\text{isp}}$. Hence, the threat that the ISP gives up caching when the CP stops paying for it is not credible anymore when the ISP already operates own caching infrastructure and in consequence Business Model 2 is a Nash equilibrium.

This change of the setting after own infrastructure is deployed is not represented in the game. It could eventually provide free caching to CPs, though, when there are not enough other paying CPs present. In consequence, the ISP cannot safely build up a working service and then offer it to the CP. Instead, the CP would have to give an incentive to the ISP to invest in caching equipment by committing to a service that is not yet working and by agreeing to a long-term payment. Although this would be preferable over the status quo, it might be unrealistic since the CPs very likely cannot or do not want to wait

for the service to be deployed after the agreement is made – or to make such a binding contract for a service of unknown real-life performance in the first place.

6.6.2 Incentives in a Repeated Game

The results of Section 6.5 suggest that Business Model 3 is not an equilibrium, since the CP has no incentive to make the payment if the ISP deploys caching. However, this outcome is Pareto-efficient if the conditions in Section 6.5.2 are met. Now, it is an interesting question whether there is a way to support Business Model 3 as an equilibrium by changing the game setting such that the CP has an incentive to pay for the caching solution after the system is set up. This would support the ISP's decision of deploying caching.

The problem with the current game model is that the ISP and the CP make their decisions independently at the same time and these are one-time choices. The game model could be improved in many ways to make it more realistic. For example, the deployment of caching services takes some time, and thus the ISP first decides whether or not to deploy caching and after the caching system is up and running, the CP decides whether or not to pay for it. This could be modeled as a two-stage game but it would not change the fact that the CP has no incentive to pay for caching.

We propose the following solution that gives an incentive to the CP. In the first stage, the ISP decides whether or not to deploy caching. If caching is deployed then they play a repeated game, shown in Table 6.4, where the CP decides whether or not to pay and the ISP decides whether or not to punish the CP (for not paying). We assume that the ISP can reduce the quality of service by QoE^- such that the CP's utility gets lower but the ISP still gets the benefits of caching. We also assume that the quality of service cannot go below the level without caching, i.e. $QoE^- \leq QoE+$, because a worse service might

not be acceptable for the CPs and also out of net neutrality considerations (although any punishment at all could turn out problematic in this regard). The other utilities are as before, e.g. $V_4 = V_2 + b_{\text{distant}} \cdot p_{\text{caching}}$.

		ISP: Punish	ISP: No Punishment
CP	No Payment	V_2 / $U_2 - QoE^-$	V_2 / U_2
	Pay	V_4 / $U_4 - QoE^-$	V_4 / U_4

Figure 6.4: Utility matrix of the punishment game.

Now, the outcome with caching and paying can be sustained as an equilibrium if the players interact for several periods and they are patient enough, i.e. they value future utilities enough. The supporting mechanism could, for example, be a simple trigger strategy. The players are supposed to choose paying and not punishing unless the other party deviates, which triggers a punishment. The ISP's punishment is to degrade the quality of service for the following periods if the CP did not pay, and similarly the CP can punish the ISP by not paying if the ISP did not offer a good caching service. We note that the CP has no incentive to make the cache payment if the payment is higher than the value of the increase in service quality. Therefore, we need to assume that $b_{\text{distant}} \cdot p_{\text{caching}} \leq QoE^+$.

Let us now calculate the required level of patience for the CP. For simplicity, we assume that the game is repeated infinitely many times and the players discount the future utilities with a discount factor δ, where $0 < \delta < 1$. Playing (Pay, No Punishment) infinitely many times is a subgame-perfect equilibrium if the players should not deviate from the path of play when a deviation

is followed by the extreme punishment, i.e. playing (No Payment, Punish) infinitely many times [2, 16, 17]. So, the path of play should give higher utility than the best possible deviation, i.e, $U_4 \geq (1-\delta) \cdot U_2 + \delta \cdot (U_2 - QoE^-)$. The right term means that utility U_2 is received one time and then $U_2 - QoE^-$ after that. From this condition, we can solve the required discount factor $\delta_{req} = (U_2 - U_4)/QoE^-$. Thus, the CP should pay for caching if $\delta \geq \delta_{req}$, i.e. if the CP is patient enough.

We note that there are also other mechanisms for supporting the (Pay for Caching, Cache) outcome of the business model game. For example, we could model the situation as a cooperative game where the ISP promises to deploy the caching system if the CP pays for it and both negotiate a suitable contract. However, this means that the ISP would have to negotiate with multiple CPs for a payment scheme, and it would be more complicated than negotiating with only one party. It might actually be more convenient for the ISP to raise a fund and promise to deploy caching if it can collect enough money from CPs, e.g. onetime payments or some usage-based contracts. But as pointed out before, these options require a huge upfront commitment by the CPs. A long-term incentive to pay for caching may also exist when several CPs compete in service quality and the ISP's caching system has a capacity too limited to serve them all: since the caches of a CP will be replaced by data of another client when the payment stops, the CP has no incentive to stop paying.

6.7 Conclusions

This chapter contributes a business perspective on ISPs in a caching environment. We showed that a market where CPs contribute to the cost of ISP caching (Business Model 3) is potentially unstable: Since the service without a caching payment is always better for the CP in the case of caching (Busi-

ness Model 2), there is an incentive to switch the contract once the caches are installed. However, Business Model 3 can be Pareto-optimal, which gives ISPs and CPs the incentive to cooperate and establish ISP caching with long-term contracts and an appropriate pricing. Further, we have shown how ISP caching can be stabilized in order to work without an upfront commitment by the CP. Our solution encourages that ISPs willfully annihilate the positive effects of ISP caching for those clients that do not pay for the system in order to encourage CPs to pay their share.

Additionally, ISP caching may use cloud services to obtain the required resources. According to our research, it is financially reasonable that providers of cloud storage choose an appropriate physical location to encourage this.

After our demonstration that ISP caching is a conclusive business model with mutual benefits to both ISP and CPs, further studies have to show whether content distribution via ISP caching is better from a technical perspective than the current practice with CDNs. For instance, a different number and location of caches could be better for different content (e.g. [21] describes an optimization regarding client latency and server capacity). Since ISPs have a higher degree of freedom in cache placement and can co-locate caching infrastructure with existing network infrastructure much easier than CDNs, ISPs do not have to compromise as much on the number and placement of caches. Also, because ISPs are in charge of routing, they might cache more efficiently than CDNs. A combination of strategic cache locations and clever routing can disperse network capacity utilization and hence keep infrastructure costs down for the ISP.

Future studies also have to address caching between networks of several ISPs. We assume in our business model game that the ISP of a CP takes care of this, for instance by peering with other ISPs (Section 6.3). While peering agreements are a possibility for cooperation between ISPs of the same size and cache utilization, compensation policies in case of unbalanced use of caches

have to be discussed. Agreements also have to be made in case that individual ISPs do not participate in caching. Service level agreements and violation penalties can help to ensure a certain quality of the service between different ISPs and between ISP and CP as well.

"The law, always behind the times, requires elaborate stitching and fitting to adapt it to this newly perceived aspect of the commons."

– Garrett Hardin

7 Discussion

7.1 Introduction

Different aspects of the IaaS cloud market have been modeled and analyzed in game-theoretic models throughout this thesis. This last chapter seeks to piece important results together in order to get the whole picture of the prospective market development. Section 7.2 derives the most likely market outcome from the research in this work. It presents key factors that are significantly involved in the market forming process. It further shows how these crucial points affect who eventually benefits from cloud infrastructure and hence may provide an opportunity for market regulation. Section 7.3 briefly addresses different market forms and Section 7.4 concludes this thesis.

7.2 A Prospect on the Future of Cloud Infrastructure

Chapter 2 pointed out that the cloud market is quite diverse, today. Despite that, there is substantial research on how and on what basis computing resources can be allocated from different available services. Recently, the trade of cloud capacities in a wholesale exchange platform was announced [27]. Such projects require comparable cloud products and similar services can be

expected to converge to the same standards in order to make them comparable.

According to Chapter 4, a market where two providers offer an identical service is unstable. Bertrand competition is an established economic model that shows that a good is sold at marginal cost in a duopoly (Chapter 4). Although the market can be stable at such a low price level, the necessary assumption of identical providers is unrealistic. Differences in size or the financial situation of the competitors affect their production cost and lead to destructive competition. Complex pricing schemes like reserved instances with an additional on-demand payment are not sufficient to significantly change the general outcome of Bertrand competition. This backs the hypothesis of an emerging monopoly. However, with a two-part cost structure where costs are different but one provider cannot provide cheaper in general, both providers can yield positive profit in an equilibrium. Two-part and three-part prices become important in that case.

Furthermore, different applications have different requirements of technical specifications and a certain variety of different service types is likely. There is research that shows that different service qualities of an otherwise identical service can coexist in a stable market [31, 80]. These different qualities may be provided by different companies and the existence of established competitors that may start to offer other service qualities could provide a threat that prevents high prices. Different services that specialize on low latencies or high memory or something similar can provide the same market segmentation.

Regarding separate processing and storage services, there are good reasons for exclusive storage and processing services. Nevertheless, only locally separated storage facilities appear to be likely for certain applications. Risk awareness or legal restrictions regarding the physical location of data may also promote the existence of separate processing facilities, though, when storage is provided from safer but otherwise less attractive locations (Scenario 2 in Chap-

ter 5). Different data center types may be owned by different companies, which would very likely prevent extortionate prices.

Additionally, future Internet technologies and business models for ISP caching could also be a reason for several smaller facilities opposed to a concentration in a single location. The storage demand of the ISP gives an incentive to the cloud providers to give up economies of scale in order to serve ISPs in locations that are appropriate for caching (Chapter 6). Lower economies of scale also reduce the chance of destructive competition as well as the market entrance barrier since a smaller provider is able to compete with a larger provider who runs more of these small facilities.

7.3 Cloud Beneficiaries in Different Market Forms

Unsurprisingly, a monopoly appears not to be a good thing from a client's perspective. In a monopoly market where instances are only rented out on-demand, though, the possibility of hybrid clouds has a huge effect on pricing (Chapter 3). When the cloud provider rents out instances for a reservation price over longer durations, the clients most likely cannot reduce their overall costs by using a hybrid cloud setup at high public cloud prices. Without this threat, the monopoly on-demand price turns out very high.

In contrast, in a duopoly (or oligopoly) market, the prices presumably are so low that the client benefit exceeds the provider benefit by far. On the other hand, the competing providers may settle with the current market shares at a high price level. They can control the market shares quite easily by an according pricing (Chapter 4). When the pricing becomes extortionate, this could require antitrust regulation.

A market situation that is advantageous for both sides, clients and providers, appears to be an oligopoly where each provider specializes in a different service (Section 7.2).

7.4 The Tragedy of the Common Cloud

The title of this thesis refers to the *Tragedy of the Commons*, a famous paper by Garrett Hardin [47]. The core statement is that it is inevitable that limited common goods are ruined by the selfish (rational) behavior of its users because an abuse by the individual creates a higher benefit to the individual than it looses due to the impairment of the good. Accordingly, games that model such situations have no inherent solution, which needs to be recognized and corrected by mutual coercion. The idea of cloud computing is to create a public utility. The actual cloud resources are not a common property, of course, and it is not my intention to stress the analogy. Nevertheless, one could consider a brisk competitive market as a common good to the clients and this market turns into a monopoly when it is ruined. There is the need to discuss whether the concept of cloud computing works for everybody on its own or if we have to make it work. This thesis is meant as a contribution to this discussion and it indicates that the market will not regulate itself if one provider becomes prevailing, but this is not inevitable since several different vendors may establish for the various reasons presented earlier.

Decisions in a cloud context may be influenced by better knowledge of the market dynamics. Although game theory claims to model real-life decisions, it is may better regarded as a tool to make these decisions. Economic research like this thesis provides knowledge that constitutes a basis for sound decisions. A similar increase of every party's benefit in the cloud market may be unrealistic in real life. But at least all involved parties can work towards a fair market

situation and rules that promote such a situation may be established by market design.

In the light of the above remarks, this work has shown that there are design decisions for the emerging common cloud standards that should be carefully considered, especially regarding an efficient combined use of resources from different sources. Although providers might promote a de facto standard that is in their interest, clients can go against this by their demand: Foresighted clients should consider the combination of separate storage and processing services and prefer on-demand services that support an operation in hybrid clouds even if such a setup is not planned at the moment. In addition, the use of services by providers other than the market leader is encouraged in order to keep several companies in the market.

Bibliography

[1] Abhishek, V., Kash, I.A., Key, P.: Fixed and Market Pricing for Cloud Services. Proc. 31st IEEE International Conference on Computer Communications Workshops (INFOCOM), pp. 157–162 (2012)

[2] Abreu, D.: On the Theory of Infinitely Repeated Games with Discounting. Econometrica 56(2), pp. 383–396 (1988)

[3] Ahlgren, B., Dannewitz, C., Imbrenda, C., Kutscher, D., Ohlman, B.: A Survey of Information-Centric Networking. IEEE Communications Magazine 50(7), pp. 26–36 (2012)

[4] Akamai: Customer Stories: MTV Networks. Website (2012), http://www.akamai.com/html/customers/testimonials/mtvnetworks.html

[5] Alger, D.: Choosing an Optimal Location for Your Data Center. Build the Best Data Center Facility for Your Business. Cisco Press (2005)

[6] Amazon Web Services: Amazon EC2 Cost Comparison Calculator. Website, http://aws.amazon.com/en/economics/

[7] Amazon Web Services: Amazon EC2 Reserved Instances. Website, http://aws.amazon.com/ec2/reserved-instances/

[8] Amazon Web Services Blog: Amazon S3 - The First Trillion Objects. Website (2012), http://aws.typepad.com/aws/2012/06/amazon-s3-the-first-trillion-objects.html

[9] An, B., Vasilakos, A.V., Lesser, V.: Evolutionary Stable Resource Pricing Strategies (Poster Paper). ACM SIGCOMM (2009)

[10] 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.: Above the Clouds: A Berkeley View of Cloud Computing. Tech. rep., EECS Department, University of California, Berkeley (2009)

[11] AT&T: Enterprise – Content Delivery. Website (2012), http://www.business.att.com/enterprise/Portfolio/content-delivery/

[12] Axelrod, R., Hamilton, W.D.: The Evolution of Cooperation. Science 211, pp. 1390–1396 (1981)

[13] Barish, G., Obraczka, K.: World Wide Web Caching: Trends and Techniques. IEEE Communications Magazine 38, pp. 178–184 (2000)

[14] Baun, C.: Untersuchung und Entwicklung von Cloud Computing-Diensten als Grundlage zur Schaffung eines Marktplatzes. Dissertation, Hamburg (2011)

[15] Baun, C., Kunze, M., Nimis, J., Tai, S.: Cloud Computing: Web-Basierte Dynamische IT-Services. Springer, Berlin (2010)

[16] Berg, K., Kitti, M.: Equilibrium Paths in Discounted Supergames. Working paper (2012)

[17] Berg, K., Kitti, M.: Computing Equilibria in Discounted 2×2 Supergames. Computational Economics 41, pp. 71–78 (2013)

[18] Born, A.: Get Off of My Cloud. iX special: Cloud, Grid, Virtualisierung, pp. 16–19 (2010)

[19] Chao, Y.: Strategic Effects of Three-Part Tariffs Under Oligopoly. International Economic Review 54(3), pp. 977–1015 (2013)

[20] Chellappa, R.: Intermediaries in Cloud-Computing. INFORMS Meeting. Talk. Dallas, Texas (1997)

[21] Chen, Y., Katz, R.H., Kubiatowicz, J.: Dynamic Replica Placement for Scalable Content Delivery. Revised Papers from the First International Workshop on Peer-to-Peer Systems (IPTPS), pp. 306–318 (2002)

[22] Cho, K., Jung, H., Lee, M., Ko, D., Kwon, T.T., Choi, Y.: How can an ISP Merge with a CDN? IEEE Communications Magazine 49(10), pp. 156–162 (2011)

[23] Christmann, C., Falkner, J., Kopperger, D., Weisbecker, A.: Schein oder Sein. iX special: Cloud, Grid, Virtualisierung, pp. 6–9 (2010)

[24] Cisco Systems: Cisco Visual Networking Index: Forecast and Methodology, 2012-2017. White Paper (2013)

[25] Cournot, A.A.: Recherches sur les Principes Mathématiques de la Thèorie des Richesses. Hachette, Paris (1838)

[26] Darwin, C.: The Descent of Man, and Selection in Relation to Sex. John Murray, London (1871)

[27] Deutsche Börse Cloud Exchange AG: The Vendor Neutral Marketplace for the Cloud. Website (2013), http://www.dbcloudexchange.com/

[28] Dikaiakos, M.D., Katsaros, D., Mehra, P., Pallis, G., Vakali, A.: Cloud Computing: Distributed Internet Computing for IT and Scientific Research. IEEE Internet Computing 13, pp. 10–13 (2009)

[29] Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R., Weihl, B.: Globally Distributed Content Delivery. IEEE Internet Computing 6(5), pp. 50–58 (2002)

[30] DiPalantino, D., Johari, R.: Traffic Engineering vs. Content Distribution: A Game Theoretic Perspective. Proc. 28th IEEE International Conference on Computer Communications Workshops (INFOCOM), pp. 540 –548 (2009)

[31] Dube, P., Jain, R., Touati, C.: An Analysis of Pricing Competition for Queued Services with Multiple Providers. ITA Workshop (2008)

[32] Dán, G.: Cache-to-Cache: Could ISPs Cooperate to Decrease Peer-to-Peer Content Distribution Costs? IEEE Transactions on Parallel and Distributed Systems 22(9), pp. 1469–1482 (2011)

[33] Eucalyptus Systems: Website, http://www.eucalyptus.com/

[34] Faratin, P.: Economics of Overlay Networks: An Industrial Organization Perspective on Network Economics (2007), Computer Science and AI Lab, MIT

[35] Fisher, R.: The Genetical Theory of Natural Selection, pp. 141 – 143. The Clarendon Press (1930)

[36] Fitchard, K.: Forget the CDN players, Netflix is Caching its own Video. Website (2012), http://gigaom.com/video/forget-the-cdn-players-netflix-is-caching-its-own-video/

[37] Flood, M.M.: Some Experimental Games. Research memorandum RM-789-1-PR, RAND Corporation, Santa-Monica, CA, USA (June 1952)

[38] Galbraith, K.: Using the Weather to Cool Data Centers. Website (2009), http://green.blogs.nytimes.com/2009/10/05/using-the-weather-to-cool-data-centers/

[39] Gartner: Gartner Says Worldwide Cloud Services Market to Surpass $109 Billion in 2012. Website (2012), http://www.gartner.com/newsroom/id/2163616;

[40] Gartner: Gartner Says Worldwide Data Center Hardware Spending on Pace to Reach $99 Billion in 2011. Website (2012), https://www.gartner.com/it/page.jsp?id=1822214

[41] Güth, W., Schmittberger, R., Schwarze, B.: An Experimental Analysis of Ultimatum Bargaining. Journal of Economic Behavior & Organization 3(4), pp. 367 – 388 (1982)

[42] Gellman, R.: Privacy in the Clouds: Risks to Privacy and Confidentiality from Cloud Computing. World Privacy Forum (2009)

[43] Google: Google Global Cache. Website (2012), http://ggcadmin.google.com/ggc

[44] Griva, K., Vettas, N.: On Two-Part Tariff Competition in a Homogeneous Product Duopoly. Tech. Rep. 9106, C.E.P.R. Discussion Papers (2012)

[45] Hamilton, J.: Cloud Computing Economies of Scale. MIX'10 Talk (2010), http://channel9.msdn.com/events/MIX/MIX10/EX01/

[46] Hamilton, W.D.: Extraordinary Sex Ratios. Science 156, pp. 477–488 (1967)

[47] Hardin, G.: The Tragedy of the Commons. Science 162, pp. 1243–1248 (1968)

[48] Hofbauer, J., Sigmund, K.: Evolutionary Game Dynamics. Bulletin of the American Mathematical Society 40, pp. 479–519 (2003)

[49] HostedFTP.com: Amazon S3 and EC2 Performance Report How fast is S3? Website (2009), http://hostedftp.wordpress.com/2009/03/02/

[50] Hustinx, P.: Data Protection and Cloud Computing under EU Law. Third European Cyber Security Awareness Day BSA, European Parliament (2010)

[51] Jalaparti, V., Nguyen, G.D., Gupta, I., Caesar, M.: Cloud Resource Allocation Games. Tech. rep., University of Illinois, Urbana-Champaign (2010)

[52] Jiang, W., Zhang-Shen, R., Rexford, J., Chiang, M.: Cooperative Content Distribution and Traffic Engineering in an ISP Network. Proc. 11th International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), pp. 239–250 (2009)

[53] Kamiyama, N., Mori, T., Kawahara, R., Harada, S., Hasegawa, H.: ISP-Operated CDN. Proc. 28th IEEE International Conference on Computer Communications Workshops (INFOCOM), pp. 49–54 (2009)

[54] Kashef, M.M., Altmann, J.: A Cost Model for Hybrid Clouds. Proc. 8th international conference on Economics of Grids, Clouds, Systems, and Services, pp. 46–60 (2012)

[55] Kaufman, L.: Data Security in the World of Cloud Computing. IEEE Security & Privacy 7, pp. 61 – 64 (2009)

[56] Khare, V., Zhang, B.: Making CDN and ISP Routings Symbiotic. Proc. 31st International Conference on Distributed Computing Systems (ICDCS), pp. 869–878 (2011)

[57] Klems, M., Nimis, J., Tai, S.: Do Clouds Compute? A Framework for Estimating the Value of Cloud Computing. Designing E-Business Systems. Markets, Services, and Networks, pp. 110–123. Springer (2009)

[58] Koomey, J.: A Simple Model for Determining True Total Cost of Ownership for Data Centers (White Paper). Uptime Institute (2007)

[59] Künsemöller, J., Brangewitz, S., Haake, C.J., Karl, H.: Provider Competition in Infrastructure-as-a-Service. Proc. 11th IEEE International Conference on Services Computing (SCC), pp. 203–210 (2014)

[60] Künsemöller, J., Karl, H.: A Game-Theoretical Approach to the Benefits of Cloud Computing. Proc. 8th International Workshop on Economics of Grids, Clouds, Systems, and Services (GECON). Springer LNCS 7150, pp. 148–160 (2011)

[61] Künsemöller, J., Karl, H.: On Local Separation of Processing and Storage in Infrastructure-as-a-Service. Proc. 9th International Conference on Economics of Grids, Clouds, Systems, and Services (GECON). Springer LNCS 7714, pp. 125–138 (2012)

[62] Künsemöller, J., Karl, H.: A Game-Theoretic Approach to the Financial Benefits of Infrastructure-as-a-Service. Future Generation Computer Systems (FGCS), pp. 44–52 (2014)

[63] Künsemöller, J., Zhang, N., Berg, K., Soares, J.: A Game-Theoretic Evaluation of ISP Business Models in Caching. Working Paper (2013)

[64] Künsemöller, J., Zhang, N., Soares, J.: ISP Business Models in Caching. Proc. 29th International Conference on Data Engineering Workshops (ICDEW), pp. 279 – 285 (2013)

[65] Lamberth, S., Weisbecker, A.: Wirtschaftlichkeitsbetrachtungen beim Einsatz von Cloud Computing. Vom Projekt zum Produkt, pp. 123–136. GI (2010)

[66] Lee, S., Jiang, J., Chiu, D.M., Lui, J.: Interaction of ISPs: Distributed Resource Allocation and Revenue Maximization. IEEE Transactions on Parallel and Distributed Systems 19(2), pp. 204–218 (2008)

[67] Level3: News Archive: Level3 Releases Statement to Clarify Issues in Comcast/Level3 Interconnection Dispute. Website (2010), http://level3.mediaroom.com/index.php?s=23600&item=65047

[68] Lewontin, R.C.: Evolution and the Theory of Games. Journal of Theoretical Biology 1, pp. 382 – 403 (1961)

[69] Liu, H.: Amazon Data Center Size. Website (2012), http://huanliu.wordpress.com/2012/03/13/amazon-data-center-size/

[70] Londoño, J., Bestavros, A., Teng, S.H.: Collocation Games and their Application to Distributed Resource Management. Tech. rep., Boston University, Computer Science Department (2009)

[71] Luce, R., Raiffa, H.: Games and Decisions: Introduction and Critical Survey, pp. 90–94. Wiley, New York (1957)

[72] Mas-Colell, A., Whinston, M., Green, J.: Microeconomic Theory, pp. 388–389. Oxford University Press New York (1995)

[73] Mell, P., Grance, T.: The NIST Definition of Cloud Computing. Tech. rep., National Institute of Standards and Technology, Information Technology Laboratory (2009)

[74] Nash, J.F.: Equilibrium Points in n-Person Games. Proc. National Academy of Sciences of the United States of America 36(1), pp. 48–49 (1950)

[75] Netflix: Netflix Open Connect Content Delivery Network. Website (2012), https://signup.netflix.com/openconnect

[76] von Neumann, J.: Zur Theorie der Gesellschaftsspiele. Mathematische Annalen 100(1), pp. 295–320 (1928)

[77] von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944)

[78] Oi, W.Y.: A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly. The Quarterly Journal of Economics 85(1), pp. 77–96 (1971)

[79] Openstack: Website, http://www.openstack.org/

[80] Pal, R., Hui, P.: Economic Models for Cloud Service Markets. Proc. 13th International Conference on Distributed Computing and Networking (ICDCN), pp. 382 – 396 (2012)

[81] Pardalos, P., Migdalas, A., Pitsoulis, L.: Pareto Optimality, Game Theory and Equilibria. Springer Optimization and Its Applications, Springer (2008)

[82] Pearson, S.: Taking Account of Privacy when Designing Cloud Computing Services. Proc. 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing (CLOUD), pp. 44–52 (2009)

[83] Peter D. Taylor and Leo B. Jonker: Evolutionary Stable Strategies and Game Dynamics. Mathematical Biosciences 40, pp. 145 – 156 (1978)

[84] Peterson, L., Davie, B.: Framework for CDN Interconnection – draft-ietf-cdni-framework-01. Website (2012), http://tools.ietf.org/html/draft-ietf-cdni-framework-01

[85] Rayburn, D.: Akamai Developing A Licensed CDN (LCDN) Offering For Telcos and Carriers. Website (2011), http://blog.streamingmedia.com/the_business_of_online_vi/2011/06/akamai-looking-to-develop-a-licensed-cdn-offering-for-telcos-and-carriers.html

[86] Rayburn, D.: CDN Pricing Stable: Survey Data Shows Pricing Down 15% This Year. Website (Sept 2012), http://blog.streamingmedia.com/2012/09/cdn-pricing-stable-survey-data-shows-pricing-down-15-this-year.html

[87] Rieck, C.: Spieltheorie. Christian Rieck Verlag, Eschborn, 6. Auflage edn. (2006)

[88] Rieck, C.: Was ist Spieltheorie? Website (2011), http://www.spieltheorie.de/Spieltheorie_Grundlagen/was-ist-spieltheorie.htm

[89] Rosenthal, R.W.: Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox. Journal of Economic Theory 25(1), pp. 92 – 100 (1981)

[90] Schmalensee, R.: Monopolistic Two-Part Pricing Arrangements. The Bell Journal of Economics 12(2), pp. 445–466 (1981)

[91] Schulte, W.R., Natis, V.Y.: "Service Oriented" Architectures, Part 1. Gartner (1996)

[92] Selten, R.: Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. Zeitschrift für die gesamte Staatswissenschaft, pp. 301–324, 667–689 (1965)

[93] Shafer, J.: I/O Virtualization Bottlenecks in Cloud Computing Today. Proc. 2nd Conference on I/O Virtualization (WIOV), p. 5 (2010)

[94] Shakkottai, S., Srikant, R.: Economics of Network Pricing with Multiple ISPs. IEEE/ACM Transactions on Networking 14(6), pp. 1233–1245 (2006)

[95] Shrimali, G., Akella, A., Mutapcic, A.: Cooperative Interdomain Traffic Engineering Using Nash Bargaining and Decomposition. IEEE/ACM Transactions on Networking 18(2), pp. 341–352 (2010)

[96] Smith, J.M., Price, G.R.: The Logic of Animal Conflict. Nature 246(5427), pp. 15–18 (1973)

[97] Smith, J.M.: Evolution and the Theory of Games. Cambridge University Press, Cambridge, UK (1982)

[98] Soares, J., Carapinha, J., Melo, M., Monteiro, R., Sargento, S.: Resource Allocation in the Network Operator's Cloud: A Virtualization Approach. Proc. 2012 IEEE Symposium on Computers and Communications (ISCC), pp. 000800 – 000805 (2012)

[99] Telefonica: Telefonica Content Delivery Network. Website (2012), http://www.telefonica.com/cdn/en/index.shtml

[100] Uptime Institute: Website, http://www.uptimeinstitute.com/

[101] Vakali, A., Pallis, G.: Content Delivery Networks: Status and Trends. IEEE Internet Computing 7(6), pp. 68–74 (2003)

[102] Vaquero, L., Rodero-Merino, L., Caceres, J., Lindner, M.: A Break in the Clouds: Towards a Cloud Definition. ACM SIGCOMM Computer Communication Review 39(1), pp. 50–55 (2009)

[103] der Veen, V.: Orange and Cogent Fight over Traffic Management before French Competition Watchdog. Website (2011), http://www.futureofcopyright.com/home/blog-post/2011/09/08/orange-and-cogent-fight-over-traffic-management-before-french-competition-watchdog.html

[104] Wang, W., Li, B., Liang, B.: Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing. Proc. IEEE 32nd International Conference on Distributed Computing Systems (ICDCS), pp. 425–434 (2012)

[105] Wei, G., Vasilakos, A., Zheng, Y., Xiong, N.: A Game-Theoretic Method of Fair Resource Allocation for Cloud Computing Services. The Journal of Supercomputing 54, pp. 252–269 (2010)

[106] Yin, X.: Two-Part Tariff Competition in Duopoly. International Journal of Industrial Organization 22(6), pp. 799–820 (2004)

[107] Zheng, X., Martin, P., Powley, W., Brohman, K.: Applying Bargaining Game Theory to Web Services Negotiation. Proc. 2010 IEEE International Conference on Services Computing (SCC), pp. 218–225 (2010)

[108] Zimmermann, H.: OSI Reference Model–The ISO Model of Architecture for Open Systems Interconnection. IEEE Transactions on Communications 28(4), pp. 425 – 432 (1980)

List of Figures

List of Tables

List of Abbreviations

CCDF	Complementary Cumulative Distribution Function	37
CDN	Content Distribution Network	129
CSP	Cloud Storage Provider	134
CP	Content Provider	129
ESS	Evolutionarily Stable Strategy	26
HPCaaS	High-Performance-Computing-as-a-Service	12
IaaS	Infrastructure-as-a-Service	2
ISP	Internet Service Provider	129
IT	Information Technology	1
NIST	National Institute for Standards and Technology	11
OSI	Open Systems Interconnection	10
PaaS	Platform-as-a-Service	11
SaaS	Software-as-a-Service	11
QoE	Quality of Experience	136
SLA	Service Level Agreement	12
SOA	Service-Oriented Architecture	8
TCO	Total Cost of Ownership	30
XaaS	Everything-as-a-Service	12

Acknowledgements

I want to thank everybody who contributed to this thesis in fruitful discussions and with review. Namely my advisors, Holger Karl and Claus-Jochen Haake, who were always up to support and advance my studies without confining the liberty of my approach. Many thanks to Sonja Brangewitz, Nan Zhang, João Soares and Kimmo Berg, who worked in close collaboration with me on the research in Chapters 4 and 6.

Further, I want to express my gratitude to the University of Paderborn and the German Research Foundation (DFG), who partially supported this work within the research training group *Automatisms* and the collaborative research center *On-The-Fly Computing*.

Warm regards to all colleagues, friends and family who accompanied me over the years of my research and who were an inspiring and helping surrounding for finishing this thesis and beyond. Thanks!