Securing the public interest in electricity generation markets

The myths of the invisible hand and the copper plate

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 29 juni 2004 om 13:00 uur

door Laurens James DE VRIES

werktuigkundig ingenieur, Master of Environmental Studies
geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. M.P.C. Weijnen

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter
Prof. M.P.C. Weijnen, Technische Universiteit Delft, promotor
Dr. ir. R.A. Hakvoort, Technische Universiteit Delft, toegevoegd promotor
Prof. W.L. Kling, Technische Universiteit Delft
Prof. B.F. Hobbs, Johns Hopkins University
Prof. I.J. Pérez-Arriaga, Universidad Pontificia Comillas de Madrid
Prof. J.M. Glachant, Université Paris XI
Prof. J. Bauer, Michigan State University

ISBN 90-5638-123-7

Cover design: Rudi Hakvoort (background: enlargement of the UCTE map of the European electricity transmission networks, courtesy of UCTE)

For Deborah

Preface

I would like to thank Prof. Weijnen for offering me a Ph.D. position in a highly inspiring environment and for her guidance in seeing me through the project. My advisor Rudi Hakvoort provided invaluable support, not only through all the long hours he spent coaching and, later, editing, but also through the opportunities he created to observe and, in a modest way, contribute to actual policy making. The Dutch Office for Energy Regulation (DTe) contributed significantly to this project through some key research assignments. I would like to thank my committee members for their guidance and constructive criticism. I am especially grateful to Prof. Hobbs for the thoroughness of his comments and his support in developing my responses. In addition, I would like to thank my colleagues Hamilcar Knops, François Boisseleau and Virendra Ajodhia for being such good team players in our electricity research group. Working together with experts from different disciplines was not only helpful and stimulating; it also made visiting conferences more fun. I would further like to thank Gijsbert Zwart for his help with the model in the Appendix. My mother, Jane, and my wife, Deborah, helped editing large parts of the final draft. Finally, I would like to thank my parents for all the support they have given me to get to the point where I could start a Ph.D. project, and Deborah, Timo and Astrid for their patient endurance of especially the last phase of this project.

LJdV

Table of contents

1 Introduction

1.1 Reason for this study

When an electricity sector is said to be liberalized, this is a simplification, as competition can only be introduced into certain parts of the electricity sector. The result is a mix of competitive activities and regulated monopoly activities. This, combined with the specific technical characteristics of electricity, is the reason that electricity markets function rather differently from other markets.

The term liberalization is also a euphemism, at least in the case of the electricity sector, as the shift from the direct regulation of a vertically integrated monopoly to the careful design of a hybrid market usually means an increase, rather than a decrease, of regulatory involvement. Given the high social and economic value of a stable supply of electricity, it is important to understand the specific dynamics of electricity markets so their design can be adjusted accordingly. This study investigates market design issues with respect to the long-term dynamics of the market for electricity generation capacity, the most capital-intensive of the liberalized functions in the electricity supply industry.

The liberalization of electricity markets is part of a broader program of liberalization of monopolies and privatization of state enterprises. The general approach to the liberalization of a hybrid sector like the electricity industry is to introduce competition where possible and to regulate the remaining natural monopoly activities. Activities that can be provided by a competitive market include electricity generation, trade, delivery to consumers (including the billing), metering of consumption and the provision of certain ancillary (support) services such as reactive power management and operating reserves. It appears that at the outset of liberalization, at least in Europe, the technical complexity of the electricity sector was underestimated. As a consequence, certain economic externalities, and the resulting possibilities of market failure, were not anticipated. This project intends to contribute to a better design of European wholesale electricity markets with respect to the policy goals of reliability and affordability of electricity. It does so by focusing upon the relationships between the technical characteristics of the electricity system and the design of generation markets.

A consequence of neglecting the technical characteristics of the system and subsequently

correcting market flaws as they become apparent is that electricity markets are not designed at once but in an evolutionary process. While this may be inevitable because the full complexity of the market could perhaps never have been understood in advance, it has some significant drawbacks. For one, over time, the compilation of *ad hoc* adjustments may lead to an overall market design that is less than optimal. A second consequence of the evolutionary approach is that some shortcomings of a specific market design (or the combination of initial design, compromises and *ad hoc* measures) may not be recognized in time, as was the case in California. The consequences may be so costly that they eclipse the potential benefits from liberalization. A third issue with the evolutionary approach to liberalization is that it creates regulatory uncertainty during the long phase until the regulations have crystallized into a more permanent form. Regulatory uncertainty can in itself be a cause of market failure as it may deter investment.

For these reasons this study steps back from the heat of the current political debates and investigates some on the long-term effects of liberalization of electricity markets. While the analysis is as general as possible, where a choice needs to be made the focus is on European electricity markets. Europe has made fundamentally different policy choices than the USA with respect to the structure of the generation market and transmission pricing. However, much of the scientific literature is focused upon American electricity systems. This project aims to contribute to an efficient and robust design of European electricity markets.

1.2 Research subject

At the outset of liberalization of the European electricity markets, the assumption was that electricity generation could be a 'normal' competitive activity, as long as the network monopoly was regulated. Regulation of the networks was deemed necessary to keep the network managers from taking monopoly rents and to ensure equal access to the networks for all market parties. The electricity generation market was expected to produce an electricity efficiently and to also invest optimally, so the future supply of electricity would continue to be optimal. The famous 'invisible hand' of the market was expected to work in the electricity market like in other markets. The electricity crisis in California cast widespread doubt upon this assumption among the general public. Some doubt already existed among experts, however, witness the presence of mechanisms to stimulate investment in generation capacity in systems such as the former England and Wales Pool, Spain, Columbia and the PJM system in the USA. The largest part of this study is dedicated to the question of whether markets can be expected to produce an optimal volume of generation capacity and, especially, what policy alternatives are available.

A second aspect of investment in electricity generation capacity is its relation to the electricity network. The most obvious aspect of this relation is the need for adequate network capacity between generators and consumers. The demand for network capacity is affected by the locations of generators and consumers. Networks were designed with a certain geographical pattern of generation and demand in mind. The new freedom for consumers to choose their generating company and vice versa means that electricity

networks now need to be able to accommodate different electricity flow scenarios than those for which they were designed. Generators may decide (or be forced) to close their doors or they may choose different locations for new facilities. However, the electricity network is not a 'copper plate', through which indiscriminate volumes of electricity can be transported between any two points, as is sometimes assumed. Decisions by generators and consumers may have a substantial impact upon the costs and even the reliability of the network.

While the physical needs for coordinating the operation of, and investment in, generation and network facilities have not disappeared, liberalization poses a significant challenge to coordination. From an economic perspective, it is important that the monopoly functions are separated – 'unbundled' – from the competitive functions in order to avoid a situation in which one market party can use control of a monopoly function to further his competitive position. If the goal of economic efficiency is to be achieved through the introduction of competition, unbundling is a prerequisite for a level playing field. This means, however, that system planning no longer is an option for ensuring efficient and effective coordination of network and generation facilities. Because the generation market is liberalized, the most attractive solution is to provide the generating companies with economically efficient incentives. This solution is hampered, however, by the existence of network externalities and the desire to create simple and transparent conditions for the generation market.

When liberalization leads to decentralized control of the system, the question is how to shape the relations between the generation market and the network so that the goals of economic efficiency and reliability are not jeopardized. The underlying hypothesis is that in a decentralized system, of which the long-term development is largely guided by economic incentives, the flaws in these incentives will cause a loss to economic efficiency and may even reduce the reliability of service.

These are the two issues that will be addressed by this study: the volume of generation capacity and how to coordinate the development of the generating stock with the electricity networks. Both issues impact primary goals of liberalization: economic efficiency and the reliability of electricity service. The central question is whether there is a need for policy intervention in European generation markets in order to meet the goals of affordability and reliability of electricity service and, if so, what policy alternatives exist and which choices need to be made.

1.3 Social and scientific relevance

The electricity sector provides a service which has become indispensable to modern society. While the turnover of the electricity supply industry is only of a few percent of the GDP of Western countries, its value becomes apparent when the electricity system fails. A conservative estimate of the cost to consumers of the crisis in the California electricity sector in 2000 and 2001 was 3.5% of California's annual economic output (Weare, 2003). Electricity is necessary for a great majority of processes in our society, as a result of which an interruption of service is an economic and social disruption of the

first order. Therefore careful *ex ante* consideration of the potential long-term effects of the current market design and proposed policy interventions is called for.

Scientifically, this project contributes to the understanding of generation market dynamics and, in particular, the merits of various policy alternatives securing generation adequacy and coordinating the development of the generation market with the networks. From a methodological perspective there is a dilemma: while scientific analysis preferably is supported by empirical data, society wishes to avoid examples of failure of electricity markets. Therefore this project conducts an *ex ante* analysis of possible market developments, which is done through a combination of qualitative analysis and modeling.

A fundamental question in this project is to what extent the physical characteristics of the electricity sector require economic arrangements that are different from 'normal' markets and that are currently not present. Thus, the relation between the physical electricity system and the electricity market is central. An interdisciplinary approach is used, which differs substantially from a purely technical or economic analysis. A monodisciplinary approach would miss the interactions between the technical and economic aspects of the system. Thus this project also contributes to the methodology for studying hybrid, infrastructure-related markets.

1.4 Reading guide

Chapter 2 provides the formal introduction to the research project: problem definition, research question, scope and assumptions. Chapter 3 provides a description of the conceptual framework, which is used in the analysis of the subsequent chapters. This framework is the lens through which the sector will be regarded in the following chapters, even if it is not always used explicitly. While originally the project focused equally upon the issues of generation adequacy and coordination of generation with the network, the more immediate nature of the question of generation adequacy, together with more rapid developments in that field, have led to a stronger emphasis on that subject.

The analysis starts with a case study of the electricity crisis in California in Chapter 4. As complex as the crisis was, and although error was piled upon error, the case study nevertheless provides useful insights into the issue of generation adequacy. These insights form the basis of the analysis of investment in generation capacity in liberalized electricity markets in Chapter 5. This chapter considers the question of whether liberalized electricity markets can be expected to provide an adequate volume of generation capacity. Interestingly, this question has not been fully addressed in the literature, even though a number of electricity systems have taken measures to stabilize investment in generation capacity. Chapter 5 concludes that there are significant reasons to doubt that liberalized electricity markets will continually produce an optimal volume of generation capacity. Chapter 6 describes a number of policy options – labeled capacity mechanisms – for stabilizing the volume of generation capacity. A framework for the evaluation of these options is developed in Chapter 7 and applied to the capacity mechanisms of Chapter 6. The analysis shows that none of the existing capacity

mechanisms are fully satisfactory for implementation in Europe, which is why some innovative solutions are explored in Chapter 8.

Chapter 9 frames the second subject of this study, the issue of coordinating the development of the generating stock with the electricity networks. This chapter provides a theoretical framework for market design issues related to this relation. One possible group of instruments for the purpose of coordination of generation with the network, congestion management methods, is analyzed in Chapter 10. This chapter develops a simple economic framework and uses this to compare several market-oriented congestion management methods. Chapter 11 provides a comparison of the two central issues, generation adequacy and coordination, and 'zooms out' to reflect upon the broader implications of the research. The conclusions and policy recommendations are summarized in Chapter 12. Figure 1.1 shows the structure of this study.

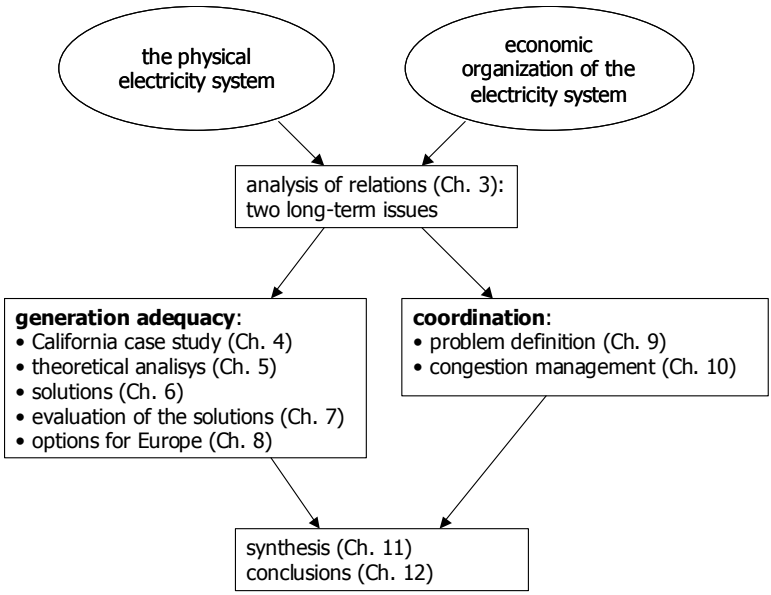


Figure 1.1: Structure of this study

Parts of this study are based upon earlier publications. The ideas for Chapter 5 were first published in De Vries and Hakvoort (2002a, 2003a and 2003b) and used ideas from Neuhoff and De Vries (2004). Chapter 7 is based upon De Vries and Hakvoort (2003c) and De Vries (2004). The latter paper also summarizes Chapter 8. De Vries et al. (2004) outlined the specific proposal of bilateral reliability contracts which was made in Chapter 8. Chapter 9 is partly based upon De Vries (2003). Chapter 10 is based upon De Vries and Hakvoort (2002b) and Knops et al. (2001).

2 Research framework

The issue of investment in generation capacity is divided into two subjects: the quantitative issue as to whether investors in competitive electricity markets produce an acceptable level of generation capacity, and the qualitative issue as to how to coordinate investment in generation capacity with the electricity networks. These issues will be addressed for current electricity systems, based on large-scale generation technology. The long life cycles and capital intensive nature of the sector mean that even a break-through of new technology will not change the fundamental characteristics of the sector immediately.

2.1 Introduction

This chapter describes the focus and scope of this research project. The next section defines the research subject, as an introduction to the research questions that are presented in Section 2.3. The research scope and assumptions are described in Section 2.4. Section 2.5 describes the research method.

2.2 Problem definition

2.2.1 Policy goals

The central question for this project is under which conditions a liberalized market in electricity generation can be expected to meet the public policy goals of economic efficiency and reliability of service. While the policy goals for the liberalization of the electricity system may vary between systems, generally economic efficiency and reliability of service receive high priority. Directive 96/92/EC, which was issued in 1996 and which formed the basis for the liberalization of electricity markets in the EU, gave as the main goal the development of a European market in order to increase the economic efficiency of the production, transmission and distribution of electricity. The ultimate purpose was to increase the competitiveness of the European economy. Secondary goals were to reinforce the reliability of service and to maintain adequate levels of environmental protection. The recent EU Directive 2003/54/EC contains a number of

measures to improve the competitiveness of the market, but also pays more attention to the reliability of electricity service, undoubtedly as a response to the electricity crisis in California.

In the USA, liberalization started earlier than in the EU with the adoption of the Public Utility Regulatory Policy Act (PURPA) in 1978, which introduced merchant generators. The act stated reliability of service as a policy goal; the goal of economic efficiency was implicit in the goal of 'equitable retail rates for consumers'.[1] The Energy Policy Act (EPAct) of 1992 further stimulated the development of an interstate wholesale market for electricity. This act also addressed some environmental issues, such as climate change and the development of renewable energy sources.

It can be concluded that both in the USA and in the EU, the general goal for restructuring the electricity system is to make it more efficient, with the purpose of reducing the cost of electricity to consumers, while maintaining or enhancing the reliability of service, within certain environmental constraints. The policy goals are not new; what is new is the means of achieving these goals, which is through competition and financial incentives rather than planning and hierarchical control. For the purpose of this project, the assumption is made that environmental policy goals are not changed with liberalization and that the same instruments remain available for meeting these goals: government sets environmental standards, which function as constraints to the electricity system. Liberalization does impact how the other two goals are met. Competition and other types of financial incentives are the new means to achieve economic efficiency. While the goals for reliability have not changed explicitly, the organizational changes brought about by liberalization have raised the issue as to which level of reliability is desired actually. The changes to the system also pose new challenges with respect to reliability and mean that new ways need to be found to maintain generation adequacy.

2.2.2 How it used to be

The electricity industry started from private initiative, and therefore competitively (Hesselmans, 1995). The natural monopoly of the networks, however, soon led to the development of regional monopolies. Governments often sanctioned these monopolies in exchange for their developing the electricity infrastructure in rural areas, which was substantially less profitable than the urban areas (Tugwell, 1988). Private utility companies eventually became regulated to curb their monopoly power. In many countries, local or national governments eventually took over and made electricity a public service (Hesselmans, 1995). A third organizational form was that of a cooperatively owned utility.

Despite these different models of ownership, the electricity companies largely functioned similarly from an economic perspective: they were all regional monopolies with a regulated revenue stream. The level of these revenues typically was based upon the expenses that the utility companies made, so there usually was no strong incentive to minimize costs. The regional monopolies had a relatively low risk-profile for investors, so capital could be obtained easily, which helped meet the particularly strong growth in

[1] PURPA (1978), 16 USC Sec. 2601.

the middle of the twentieth century.

Many electric utilities were vertically integrated, meaning that they managed the full supply chain of generation, transmission and distribution. This facilitated the planning and operation of this technically complex industry. The networks planned together with the generation facilities; generators were operated not only with a view to consumer demand, but also with consideration for network constraints and the need for voltage control. Throughout the twentieth century, increasing economies of scale dominated the development of generation facilities. These ever-larger plants were usually constructed away from urban areas, creating the need for high-voltage transmission networks. Increasingly, transmission networks were linked so the connected electricity systems could provide emergency assistance to each other. The growing scale of the interconnected network, which eventually became continent-wide, required increasingly sophisticated system planning. Eventually, the development of generation facilities and networks was planned at the level of states or countries. Investment decisions were made from a system perspective, balancing the costs and benefits of generation versus network expansion. An important advantage of central planning and operation was the ability to meet the need for coordination of network and generation operation and investment. It provided the simplest possible way of managing this technically complex industry.

The advantages of the monopoly model were that it was a convenient way of bringing electricity to rural areas, of financing a rapid rate of expansion and that it was a simple way to manage the technical complexity of the sector. The main disadvantage was considered to be the lack of sufficient incentives for economic efficiency.

2.2.3 Definitions

Before continuing, it is useful to introduce the definitions that the sector uses for issues related to reliability (UCTE, 2002b):

> **Reliability** – a general term encompassing all the measures of the ability of the system, generally given as numerical indices, to deliver electricity to all points of utilization within acceptable standards and in the amounts desired. Power system reliability (comprising generation and transmission facilities) can be described by two basic and functional attributes: adequacy and security.
>
> **Adequacy** – a measure of the ability of the power system to supply the aggregate electric power and energy requirements of the customers within component ratings and voltage limits, taking into account planned and unplanned outages of system components. Adequacy measures the capability of the power system to supply the load in all the steady states in which the power system may exist.
>
> **Security** – a measure of power system ability to withstand sudden disturbances such as electric short circuits or unanticipated losses of system components together with operating constraints. Another aspect of security is system integrity, which is the ability to maintain interconnected operations. Integrity relates to the preservation of interconnected system operation, or the avoidance

of uncontrolled separation, in the presence of specified severe disturbances.

The issue of security will not be considered, as it is outside the scope of this project, which is only concerned with the long-term development of the generation market.

In theory, these concepts can be quantified and measured. In reality, however, it is difficult to obtain some of the necessary data. Reliability is usually measured as a function of the frequency and duration of service interruptions. The reliability of the existing system can therefore be quantified without too many complications. The issue is, however, how to forecast the future reliability of the system. To ascertain that a certain reliability standard will be met in the future, it is necessary to forecast reliability at least as long ahead as it takes to realize new generation or network facilities. Only then is it possible to determine the need for new investment at each point in time. The long lead times for investments in generation capacity and for network components means that these forecasts need to be made a number of years ahead.

With respect to generation adequacy, two categories of data are required: data about the volume of available generation capacity and data regarding demand. Comparing the two, the probability can be determined as to whether the available generation capacity is sufficient to meet demand. Forecasting the demand for electricity is at least as difficult as forecasting general economic growth, as this is one of its main drivers. Before liberalization, this was the main difficulty in system planning. Liberalization, however, has added significantly to the uncertainty because now each generating company needs to make its own investment decisions. These do not only depend upon the overall demand for electricity, but also upon the development of the company's market share. In addition, the capabilities and costs of the network no longer necessarily factor into investment decisions for generation facilities. Impractical decisions regarding new generators or closure of existing generators may not only bring about extra costs for the network, but also reduce the reliability of the system.

For competitive reasons, generating companies have no interest in divulging their plans for opening or closing generation facilities. Moreover, even if they would consent to confidentially notifying an independent agent of their plans, there would be no means of binding the companies to these plans. Companies would need to retain the right to deviate from their plans to open or close plants if conditions changed (in the market, but also in their own financial situation). This means that a certain amount of uncertainty is inevitable. Finally, information regarding planned changes in generation capacity may become a tool for gaming, for instance by over-stating expansion plans to scare away competitors from investing. This would reduce the value of the information collected this way. These basic uncertainties make it quite difficult to make firm statements about the future adequacy of generation capacity or about system reliability in a liberalized market.

2.2.4 Two aspects of market design

New structure, new challenges

Before liberalization, the electricity system was a complex, technical system that was optimized in a centralized fashion. Now, it is a complex technical system with on top of it

a complex economic market structure that directs the technical system. The result is a significant increase in the complexity of the system as a whole. The different actors that make up the system optimize their own parts of the system with respect to their own goals; the performance of the system as a whole therefore depends upon the incentives that are provided to these actors.

Competing parties in a market are assumed to act rationally in their own interest, which they often do. In theory, this should not only lead to the maximization of their own utility (which usually is operationalized as wealth) but also to the greatest benefit for society. However, this is only the case when the market functions perfectly. One of the conditions for perfect competition is that there are no external benefits or costs: when each competing party bears the full costs of his actions and receives the full benefits of his actions. Electricity markets, however, are rife with externalities, both positive and negative, which are largely caused by the network monopoly. For example, an available generator that is inactive contributes positively to the reliability of service of all consumers connected to the same network but is not necessarily remunerated for this service. One of the main goals of designing a market is to minimize these externalities through the creation of efficient financial incentives. The idea is that through a careful design of the market, the benefits of competition can be maintained while the negative effects of externalities are minimized. This way, economic theory suggests, the system will tend to gravitate towards a socially optimal equilibrium state.

The fact that the electricity network has a natural monopoly means that it needs to be regulated. Network managers must be stimulated to make operational and investment decisions that correspond to the goals of economic efficiency, reliability and minimization of environmental harm. This is the first task in designing a liberalized electricity system. The generation market, on the other hand, is intended to be competitive, which means that the starting point is a minimum of regulation. This study will show, however, that additional regulation is necessary to obtain adequate investment in generation facilities in a competitive market. This is the second task in designing an electricity system. Finally, the technically close relationship between generation and the networks means that reliability is affected by the degree of coordination between the two, and that there are economies of coordination as well. Coordinating the generation market with the network monopoly is the third challenge in designing an electricity system. This project leaves aside the question of network regulation and focuses on the latter two issues.

Generation adequacy

The question of generation adequacy in a competitive market is how to achieve an optimal level of generation capacity over time. 100% reliability is technically not possible, which means that a balance must be found between the costs of improving reliability and the demand by customers for reliable service. This study focuses on generation adequacy as the long-term component of reliability. For the case of generation, the above definition of adequacy will be interpreted as the extent to which the system provides the right volume of generation capacity.

The reason that this issue is addressed is that there are indications that electricity markets

have different dynamics than other markets, as a result of which the mechanisms that balance supply and demand (the 'invisible hand') may not function as well. For instance, electricity cannot be stored in existing networks, which means that there is a need for peaking generators that operate rarely. The question is whether a market can finance these units reliably. The electricity crisis in California, which will be analyzed in Chapter 4, indicated that markets may not always provide sufficient generation capacity.

Coordination

Electricity markets do not develop naturally. If left unregulated, network managers can expand their monopoly to include generation and the delivery of electricity to consumers, so the industry would become vertically integrated. The solution is to regulate the sector and to 'unbundle' the monopoly functions from the competitive functions. This means that the parties who control the network and other monopoly activities are not allowed to be involved in competitive activities. While the need for unbundling is recognized widely (cf. Newbery, 2001; FERC 2002b; Directive 2003/54/EC), the degree to which it is applied varies.

Much attention has been given to the need for unbundling, how to implement it and how to regulate unbundled network companies. Less consideration has been given to the technical interdependencies between the networks and generation. This interdependency, which was one of the arguments why the electricity industry as a whole should be considered as a natural monopoly, still exist. There is a need to structure the relationships between the generation market and the networks efficiently, lest the economies of coordination be lost.

2.3 Research questions

From the problem definition the main research question is distilled:

Does the current design of European wholesale electricity markets provide adequate long-term incentives with respect to the goals of reliability and economic efficiency, and if not, what are the policy options?

As explained above, the two main aspects are generation adequacy and the coordination of generation investment with network development. The research question can be specified for these two subjects:

> **Generation adequacy**: does an unregulated liberalized electricity market tend to produce sufficient generation capacity over time; if not, what policy options are there for securing a sufficient volume of generation capacity?
>
> **Coordination**: is there a need for coordinating investment in generation capacity with the networks? If so, what are the policy options?

2.4 Research scope and assumptions

2.4.1 Generation adequacy

The reliability of electricity service is the product of a chain of activities. This project focuses only on one link, electricity generation. Therefore all statements referring to reliability are limited to the impact of generation capacity upon reliability. This is a significant limitation, as most regular disturbances of electricity service are a result of network failure, while in the very long term the supply of primary fuels is arguably the main concern. However, a chain is as weak as its weakest link, and overlooking the issue of generation capacity may prove a costly mistake.

The economic model of the sector that is used as a starting point for the analysis in Chapter 5 is that of an energy-only market, which is defined as a market in which the price for electric energy is the only source of revenue for recovering investments in generation capacity. This can be considered the most deregulated type of electricity market, as there are no rules concerning the structure of the market. Fundamental to the issue of generation adequacy is that actual electricity markets are rather different from the theoretical ideal. In particular, consumers typically are insufficiently involved in the market, so that demand price elasticity is extremely low. A common approach in the scientific literature is to call for improvements of the electricity market infrastructure, in particular the installation of real-time meters, so that it better resembles the theoretical ideal. These adjustments may not be forthcoming, however, considering the many ways in which reality deviates from the theoretical ideal. (See Chapter 5.) This study has a less idealistic starting point. Rather than proscribing how the technology should be improved, how governments should stop creating regulatory uncertainty and how consumers should improve their response to real-time electricity prices, et cetera, this study takes the current imperfections of electricity markets as a given and poses the question how the policy goals of economic efficiency and, in particular, reliability of service can be obtained despite these imperfections. The result is advice as to how to stabilize the volume of generation capacity in the long and difficult transition period to a stable, competitive market – and perhaps thereafter, if the market imperfections are not sufficiently removed.

While this project focuses on long-term issues (investment), an exception is made for the subject of strategic manipulation of the availability of generation capacity. The analysis in Chapter 4 shows that this can be a fundamental threat to reliability and that the incentives for manipulation are influenced by the market design. Therefore this operational aspect will be included in the analysis. The focus is upon the structural availability of generation capacity; other operational issues such as maintenance and system operation are outside the scope of this project. An extensive survey of the operational reliability in the California crisis is provided by Roe et al. (2002).

The analysis focuses upon electricity systems that are served by large-scale generation plants. This means that new generation facilities need considerable time to be realized and the generation market cannot react quickly to sudden shortages. The analysis is restricted to electricity systems in which hydropower does not play a significant role.

This is a significant assumption, as it means that the total volume of generation capacity determines the reliability of the supply of electricity. In a hydropower system, there usually is abundant generation capacity and the reliability of electricity generation is determined by the energy content of the hydro reservoirs. In other words, shortages in hydro systems usually are not the result of limited generation capacity, but of a lack of water in the system. As a result, the dynamics of a hydro system are different from a system without hydropower.

While the question of generation adequacy has not received broad attention, especially not before the California electricity crisis, the literature does provide a basic analysis of the problem (most notably by Meseguer and Pérez-Arriaga, 1997; Hobbs et al., 2001c; Doorman, 2000), and some proposals for solutions (Doorman, 2000; Vázquez et al., 2002; PJM's system of installed capacity requirements). Chapter 5 builds upon the existing literature to develop a cohesive argument as to why competitive energy-only markets would invest less in generation than optimal. Chapters 6 through 8 discuss solutions and present a policy framework.

2.4.2 Coordination

Starting point for the analysis of the coordination issue is the European model for transmission tariffs. While the details vary among the different countries in Europe, the principle is that users of electricity networks pay a fixed transmission tariff that is independent from the distance over which the electricity is transmitted. (Only when congestion occurs is this principle abandoned.) Fixed transmission tariffs provide simplicity and transparency to the electricity market, which are much-needed qualities. However, the lack of efficient economic incentives raises coordination issues. Using fixed transmission tariffs is fundamentally different from a system of locational marginal pricing, which, among others, is used in several systems in the USA. In this system, the transmission tariffs vary continuously in order to include the costs of congestion and network losses in the optimal dispatch calculations.

Again, the analysis is based upon current technology, in particular large generation plants in a system with little or no hydro power. As hydro plants are geographically fixed, there is no question about their future locations and the network can be safely developed upon the assumption that they will remain active. The assumption of large plants is significant to the extent that they do not necessarily locate near demand. Changes in their locations may therefore impact the flow of electricity through the network significantly. A transition to distributed generation would probably reduce the demand for transmission capacity, as more electricity would be generated close to consumers. It would also reduce the fluctuations in load flows, as the changes in output of many small units would largely cancel each other out (except perhaps to the degree that electricity is generated from wind and solar energy). Another technical innovation that could influence the issue would be the wide-spread introduction of power electronics, which would allow better operational control of the network. However, the long life cycle of the technical components of the electricity sector – both generators and network components – means that even if these technologies break through, it may take decades before their application has become ubiquitous. Consequently, the analysis in this study will be pertinent until that time.

2.4.3 Technical developments

This study considers the dynamics of the current electricity system, given current technology. The capital-intensive nature of the electricity sector means that any change in technology is not likely to change the system dynamics overnight, so this is a safe assumption for the foreseeable future. Nevertheless, it is useful to keep in mind that there are some technical developments that may change the dynamics of the sector, and therefore also the problem analysis that is presented in this study. The main developments are presented here briefly.

Storage

One of the physical characteristics that makes the electricity market so different from other markets is the requirement that supply and demand must match each other from moment to moment. This means that available generation capacity must always be at least as large as peak demand. It also requires great versatility on the part of generators. Commercial availability of storage technology would change the dynamics of the electricity market substantially. It would allow a smaller volume of installed generation capacity, resulting in potentially large cost savings, and remove the need to balance supply and demand continuously, facilitating system operation and reducing price volatility. Currently the only available technology is hydropower, but this is limited to mountainous regions. A new technology based upon fuel cells may prove more generally applicable, but is currently still in the pilot phase (Regenesys Technologies, 2003).

Renewable energy sources

Much emphasis is being placed upon the development of renewable sources of energy, although their current market share is small. Strong growth of the proportion of electricity generated from renewable energy sources could change the system dynamics. In principle, for our analysis the primary energy source for electricity generation makes no difference. However, some renewable sources of energy are not continuously available (most notable wind and solar energy). This complicates network operation as well as the analysis of generation adequacy. A second aspect of renewables is that they often have a dispersed nature, meaning that they provide many small sources of power, rather than large concentrated ones. As small generation units typically are linked to the distribution networks, this changes the dynamics of network operation.

Distributed generation

Until the 1980s, the optimal scale of power generation plants increased continuously. A different generating technology reversed this trend: combined-cycle gas turbines could suddenly produce electricity more cheaply at a much smaller scale. This prompted large electricity consumers to demand to be allowed to generate their own electricity, which contributed to the call for liberalization of the electricity sector (Hunt and Shuttleworth, 1996). The development of small cogeneration technology continues, which opens a perspective of an electricity system consisting of many small units, connected to the distribution network or directly at the consumers. The concept of small electric power generation facilities that are directly connected to the distribution network or to

customers is often referred to as distributed generation (Ackermann et al., 2001). Such a system would require a smaller volume of transmission capacity, as the distributed generation units would feed directly into distribution networks. Feeding this perspective is the promise of the fuel cell as an even more efficient means of generating electricity. The use of distributed sources of renewable energy such as wind and solar energy also fits into this perspective.

On the other hand is the same combined-cycle gas technology being applied at an increasingly large scale, so the large, central power plants also continue to become more efficient. As a result, it is as yet unclear whether distributed generation will become the new paradigm, or whether it will remain a niche market. There are significant benefits to be obtained with the use of distributed generation. For instance, network losses would be lower and more waste heat could be utilized (as it becomes available close to consumers). However, there also are significant obstacles, such as the large sunk costs in the existing system, which create a strong path dependency as well as the need for substantial adjustments to distribution networks in order to accommodate the changed dynamics of such a system.

Network technology

The possibilities for operational control of electricity networks are limited. Switches are relatively slow and costly to use. As a result, the flow of power through an electricity network is largely determined by the laws of nature. This leaves adjustments of generators as the main control option. In the vertically integrated utilities before liberalization, the physical limits of the networks were taken into consideration when dispatching the generators. Now that generation is unbundled from the networks in many liberalized systems, operational control of the network has become more difficult. (See Chapters 9 and 10.)

The introduction of FACTS (Flexible AC Transmission Systems) may greatly facilitate network operation (Moore and Ashmole, 1995; Moore and Ashmole, 1996; Moore and Ashmole, 1997; Moore and Ashmole 1998). The term FACTS refers to a wide category of technologies, some of which are based upon power electronics. While some applications have been in use for more than a decade, the more advanced technologies, which would allow better operational control of electricity networks, are still being developed or are not yet cost-efficient.

Electricity meters

The electricity consumption of most consumers is measured infrequently, for instance once per year. As a result, bills do not reflect the price differences of consumption at different times: consumers pay the average price for electricity during the metering period. A crude improvement is provided by double meters, which measure consumption during peak hours separately from off-peak hours. However, these still do not signal consumers when prices are higher than usual. As a result, consumers with these meters have no incentive to respond to temporary electricity shortages. Only if average prices rise for a long period of time do they lead to a noticeable increase in consumer bills which may eventually lead to an adjustment of demand. In the short term, current

electricity meters cause demand price-elasticity to be nearly absent.

Electronic meters could change this by measuring consumption per time interval. If the price paid by consumers is based upon the spot price of electricity, they will have an incentive to exhibit more price-elastic behavior. This would improve the overall economic efficiency of the system, as it would lead to lower peak consumption. However, installing these meters for every customer is a large operation which most systems have not (yet) undertaken.

2.5 Method

This study considers two aspects of the long-term dynamics of liberalized markets for electricity generation. The two issues of investment in generation capacity, adequacy and coordination, have not manifested themselves widely in practice. There is a lack of empirical evidence regarding both issues, which can be explained by the fact that the life cycle of generation facilities far exceeds the history of most liberalized electricity markets. Therefore, the approach used here is to analyze the dynamics of the electricity sector in order to determine the possible development paths for the system.

The long-term development of generation markets is a relatively little-explored subject. This may be due to the fact that in many systems more pressing issues dominated the debate during the first years following liberalization. Or perhaps it simply is a result of the presumption that investment decisions can safely be left to the market. Another reason may be that much research is mono-disciplinary, while a good understanding of the dynamics of the generation market and its relation to the network requires a combination of technical and economic analysis. This study uses a multidisciplinary approach, combining a systematic comparison of the technical characteristics of electricity systems with the economic structure.

The technical requirements of the system are compared to the structure of the electricity market with respect to the long-term incentives that it provides to the generation market. The incentives are evaluated with respect to how well they can be expected to guide the electricity system towards economic efficiency and reliability. This approach reveals which characteristics cause the electricity system to be different from other infrastructure sectors and what the consequences are for the design of the structure of a liberalized electricity sector. The conceptual framework in Chapter 3 elaborates the model that underlies this approach.

2.5.1 Generation adequacy

Empirical material

There are some examples of electricity markets that failed to produce an adequate volume of generation capacity but the case material is convoluted by non-market factors. An example is the reluctance of the Norwegian government to permit a natural gas-fired generation plant (currently nearly all electricity is generated from hydropower), which

contributed to the scarcity of electricity in the winter of 2002-2003. Brazil also faced tight supplies in recent years but again the involvement of government in the generation sector limits conclusions about possible market failure. Moreover, both Brazil and Norway are almost completely dependent upon hydropower. This is also the case in New Zealand which experienced shortages in 2003. The presence of a significant proportion of hydro power changes the dynamics of investment in generation capacity, as it is energy-constrained, rather than capacity-constrained. This means that reliability is not only determined by the volume of available peaking capacity but also by the total energy content of the reservoirs. This makes these cases less representative for systems with a limited share of hydropower. Recently, the UCTE issued a warning, based on the extensive data that it collects each year, that generation adequacy may be threatened in Europe by the end of this decade if there will not be substantial investments made in generation capacity in the near future (UCTE, 2003).

The most notable case of electricity shortages in a market-based system, California, is described in Chapter 4. The California case was chosen because it is a capacity-constrained system like most other electricity markets. Hydro power did play a role in California (mainly through imports), but only for a small portion of its generation capacity. In addition, the world-wide attention which the electricity crisis in California attracted made this case a reference point for public policy so it cannot be ignored when addressing the issue of generation adequacy. The strong attention for this case gave rise to many misconceptions leading to the need to separate fact from myth before the appropriate lessons can be learned.

The lack of case material regarding the adequacy issue is also caused by the fact that several of the forerunners of liberalization, such as the PJM system in the USA and the England and Wales Pool, had specific systems to ensure investment in generation capacity. Therefore, they provide no information about the performance of energy-only markets. From a social perspective, this situation of paucity of empirical material must be continued as the social cost of producing evidence of failure of the electricity system is extremely high – witness the cost of the California electricity crisis. Scientifically, this means that rather than building theory from empirical evidence, theory needs to be developed from the analysis of the electricity system's characteristics and the market structure.

Quantitative modeling

Another approach would be to develop a quantitative model with which system development can be forecast in order to determine whether the current system design can be expected to produce satisfactory long-term results. However, a model can never produce forecasts with full certainty. The value of model forecasts is inevitably limited by the many non-quantifiable factors that need to be incorporated, such as the impact of regulatory risk upon investment behavior, strategic behavior in an oligopolistic market, investor risk aversion and the impact of imperfect information regarding the stochastic distribution of demand and the availability of competitors' generators. Finally, the value of the forecasts made with such a model would be reduced by impending changes to market rules, such as the implementation of the new electricity directive in the EU and the proposed Standard Market Design in the USA (Directive 2003/54/EC; FERC, 2002b).

For these reasons, a deliberate choice has been made for a qualitative analysis of the question of generation adequacy rather than a modeling approach. A quantitative analysis of the generation adequacy in the Netherlands is currently being made, however, at this university, also as part of their doctoral research, by Rödel and Van Eck (Van Eck et al., 2002).

The multi-criteria analysis of capacity mechanisms in Chapters 7 and 8, on the other hand, is supported by a dynamic model (in the Appendix). While the model necessarily describes a simplified case, it allows comparison of the behavior of different market models under the same circumstances.

Approach

Starting point for the analysis of generation adequacy is the case study of California in Chapter 4. Chapter 5 steps back from practice to consider how an energy-only market should work in theory – this is, in neo-classical economic theory. Using the indications for possible causes of market failure that the analysis of the California electricity crisis provided, the chapter continues with a systematic evaluation of the reasons why there would or would not be sufficient investment in an energy-only electricity market. Both a static equilibrium and dynamic development are considered. The effects of imperfect competition are also included. While this qualitative analysis cannot predict the future development of a specific market or quantify the risk of electricity shortages, it does present a number of arguments for changing the structure of energy-only markets. The analysis of possible causes of market failure leads to a set of criteria which adjustments to the market structure should ideally meet.

Chapter 6 describes the main policy options that have been tried or proposed in the literature for the stabilization of generation capacity. These are evaluated with the policy framework that is developed in Chapter 7 and 8. The analysis in these chapter is supported by the model of the Appendix. While descriptions of these instruments can be found in the literature, a systematic comparison based upon criteria derived from an analysis of the problem is new.

2.5.2 Coordination

Empirical material

A similar lack of empirical information exists with respect to the issue of coordination. Again, the relatively short history of liberalized markets means that no significant changes in the locations of generators can be expected to have taken place. A lack of coordination may also not become apparent as quickly as a shortage of generation capacity for several reasons. One, the magnitude of the issue depends upon the average behavior of all generators, whereas the adequacy issue is one of investment at the margin. Two, the long-term costs of a lack of coordination may be limited to higher network costs (e.g. higher energy losses or congestion costs), rather than something as dramatic as service interruptions.

Approach

The issue of coordination is also addressed from the neo-classical perspective that investment should be socially optimal if the incentives are economically efficient. Here the problem is not so much the limitation of neo-classical theory but the fact that the ideal of efficient incentives appears unfeasible, at least in Europe, where the choice has been made for *ex ante* fixed transmission tariffs.

The greatest obstacle to answering the question of whether there is a need for coordination is that the answer depends strongly upon the network in question. It was already mentioned that the issue is not likely to emerge in hydropower-based systems as there generators are geographically bound. In other systems, there also may be geographical limitations to the locations of new generators. In addition, the structure of the network determines how sensitive it is to large shifts in the locations of generators. A finely meshed transmission network with a large capacity is more robust in this respect than a system with limited transmission capacity. Thus, the cost of a lack of coordination depends upon the existing system and the range of development scenarios. The fewer options there are for generators, the smaller the need for coordination as the development of the generation market is more predictable.

As little research has been done on the subject, the first step is a comprehensive inventory and structuring of the issue and possible solution paths. This is done in Chapter 9. Again, the technical characteristics of the system – the physical requirements for coordination of generation with the networks – are the starting point. Next, the options for structuring economic relations that reflect these requirements are assessed. Chapter 10 starts by investigating one category of possible solutions, namely congestion management methods.

3 System description

Before starting the analysis, this chapter describes the 'worldview' that underlies it. At every step in the analysis, the starting point is the physical electricity system. Economic efficiency means mainly that the physical system is operated and developed efficiently. To achieve this goal, the economic 'subsystem' of the electricity system, which overlays the technical 'subsystem', must provide the right signals for operation and investment. The focus of this study is the relationship between the economic and the technical subsystems of the electricity sector. This chapter develops a conceptual framework for describing these relationships.

3.1 Introduction

The conceptual framework that underlies the analyses in this study is developed in this chapter. The premise that underlies the analysis of both generation adequacy and the coordination issue is that the technical characteristics of the electricity sector cannot be ignored in the design of the market. This means that there is no standard liberalization model that can be applied to the design of electricity markets. This chapter develops a framework for analyzing the relations between the technical and economic sides of electricity systems. Even though the graphic representation does not return in each chapter, the concept of separating the technical and economic parts of the system is the basis for the analysis. In addition, the last section of this chapter discusses the recurring issue of system optimization.

3.2 The electricity system

To begin with, a few definitions need to be made. The term *electricity system* is used to indicate the combination of the systems that produce, transport and deliver power and provide related services. It includes the parties that trade in electricity or provide trade-related services such as electricity exchanges and brokerage services. The electricity system can be divided into two subsystems: a technical subsystem, centered around the production and transmission of electricity, and an economic subsystem, in which electricity and transmission services are traded. Both subsystems are constrained by

regulations, physical factors and the historical development path of the system. Figure 3.1 presents a graphic representation of this basic conceptual framework.

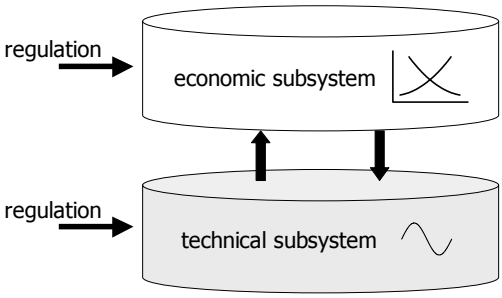


Figure 3.1: Basic framework for the electricity system

3.3 The technical subsystem

The technical subsystem consists of the hardware that physically produces and transports electric energy to customers, as well as the equipment that uses the electricity. It further consists of the people and organizations that build, maintain and operate the equipment. The structure of the technical subsystem is determined by the nature of the components that make of the electricity supply system: the generators, the transmission network, the distribution networks and the loads. This section will briefly describe the components and operation of the technical subsystem. See also the text box with definitions on page 23.

3.3.1 Components

Generators are apparatus that produce electricity from other forms of energy. Their most important characteristics are:

- size (capacity),
- controllability (speed with which they can react to changes in demand),
- availability (scheduled and unscheduled outages),
- reactive power generation capacity,
- energy source (coal, nuclear energy, wind, *et cetera*), and
- environmental impact (emissions, waste, noise).

The first four characteristics are essential to determine a generator's behavior in the transmission network. All characteristics have value in the economic subsystem. The latter two characteristics may be important for the economic subsystem, if customers demand electricity from a source that is less harmful to the environment or sustainable.

Electric energy is transmitted from generators to consumers through a highly meshed network. This network actually exists of a number of linked networks of different voltages, linked by transformers. The energy losses from transmission are less at higher voltages, which is why they are used for transmission over longer distances. Usually, the

electricity networks are divided into transmission and distribution networks; the boundary between the two is somewhat arbitrary. For the purpose of this research, the electricity networks from generator to consumer can often be considered as a whole, which is why often will be referred to 'the network'.

Definitions

Electricity system: the combination of systems that produce, transport and deliver power and provide related services, including the actors and institutions that control the physical components of the system. The electricity system consists of a technical and an economic subsystem.

Economic subsystem: the actors that are involved in the production, trade or consumption of electricity, in supporting activities or their regulation, and their mutual relations.

Technical subsystem: the physical part of the electricity system, consisting of the hardware that physically produces and transports electric energy to customers, as well as the apparatus that use the electricity.

Generator: an apparatus that produces electricity from another form of energy. Primary energy sources can be hydrocarbons, nuclear energy, or sustainable energy sources such as wind, the sun, geothermal energy and biomass. Secondary energy sources such as diesel oil or hydrogen gas may also be used.

Load: apparatus that uses electricity from the electricity network, varying from consumer appliances to industrial processes.

Line, link: terms used synonymously to indicate the links in an electricity transmission network. The main technical characteristics of electricity wires are their capacity, which is the amount of energy they can transmit, and their impedance, which is the combination of their electric resistance and phase-shifting properties.

Transmission and distribution: both terms refer to the transport of electricity. Transmission typically indicates longer distances, for which higher voltages are used, while distribution indicates local transport to end users. The transmission and distribution systems are networks. They often have multiple routes between two points to enhance system reliability. As a result, not line capacity but network capacity is the determining factor.

Transformer: apparatus that converts electricity from one voltage level to another voltage level. They are an essential part of any large-scale electricity network, as electricity is transported at high voltage levels and mostly used at much lower voltage levels.

Dispatch: operating instructions for generators.

Control zone: contiguous part of a network within which the energy balance and power quality are controlled.

Ancillary services: compensation for power losses, management of reactive power, and voltage and frequency support.

The geographical boundaries of a network are, in principle, arbitrary. Their shape is historically developed; their boundaries often coincide with political boundaries. Neighboring networks usually are linked. Here, a single network will be considered to be that part of the interconnected network that is administered by one system operator.

The dominant characteristics of transmission and distribution networks are:
- the morphology of the network, including links to networks of other voltages,
- the transmission capacity of each link,
- the impedance of each link, and
- possibilities to control voltage and reactive power.

The last large category of physical components are the loads. Loads are the apparatus of consumers that use the electricity – for lighting, power, et cetera. The most important characteristics of loads are:
- maximum demand,
- reactive power demand,
- demand pattern, and
- interruptability (can they be switched off).

3.3.2 Operation

The combined characteristics of generators, loads and the network determine how much electricity is generated and consumed. Different combinations of supply and demand result in different load patterns of the network. Two functions are needed to manage the technical subsystem, namely a system operator and a transmission operator. Although the actors that perform these functions are not themselves part of the technical subsystem (according to the definition used here) they are included in the figures, as their role is central to the functioning of the system. The system operator (SO) maintains system stability and manages the energy balance within a 'control zone', as the network itself cannot store electricity. Where actual demand and supply deviate from the amounts that were contracted by market parties, the system operator maintains the power balance continuously. If the market projected demand well, the need for balancing is small, but it is crucial for system stability. A second task is to provide (or contract) sufficient black-start power. There is one system operator per control zone.

The transmission operator (TO) manages a transmission system. He guards against congestion, maintains reliability of transmission service and provides ancillary services for transport. There can be multiple transmission operators per control zone. A similar function exists for distribution networks. The tasks of TO and SO may be joined in one agency, the transmission system operator (TSO). This is the case in European markets. Figure 3.2 shows a model of the technical subsystem. In this figure, the term network is used to indicate both transmission and distribution networks.

In order to be able to combine the technical and economic subsystems in a single model that shows the relationships between the two, the model of the technical subsystem can be simplified to the one shown in Figure 3.3. Here, generators, networks and loads are all aggregated.

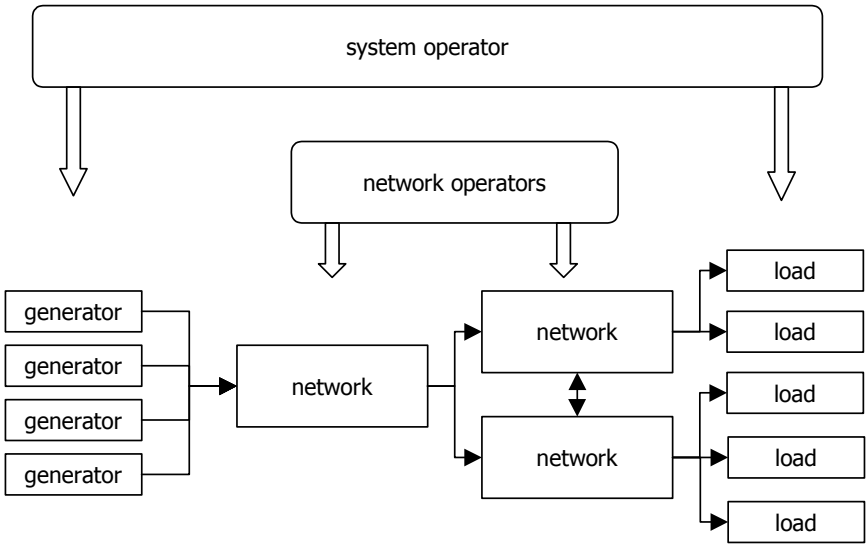


Figure 3.2: The technical subsystem

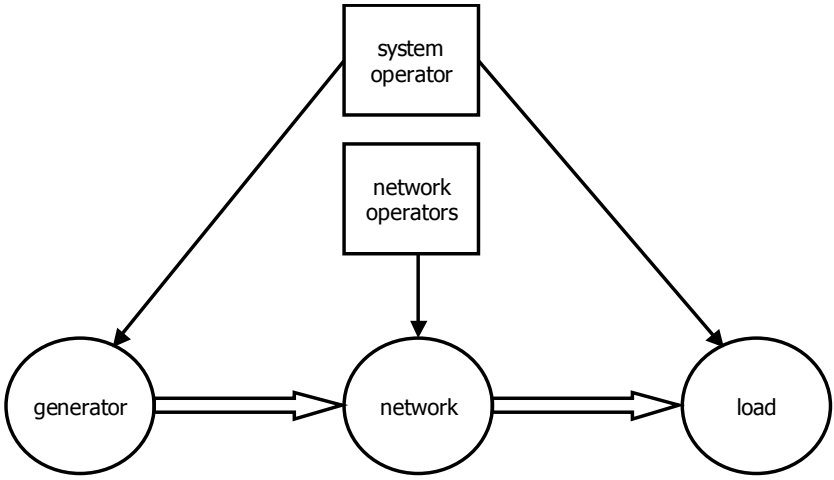


Figure 3.3: The technical subsystem, simplified

The electricity system before liberalization

Before liberalization, the electricity system consisted of little more than the technical system. (See Figure 3.4.) The utility companies owned generation facilities as well as the networks. Sometimes they controlled the entire production chain from generation to retail, as is shown in the figure; in other cases, different companies provide generation or distribution, for instance. Key is, however, that all services were provided by regulated monopolies.

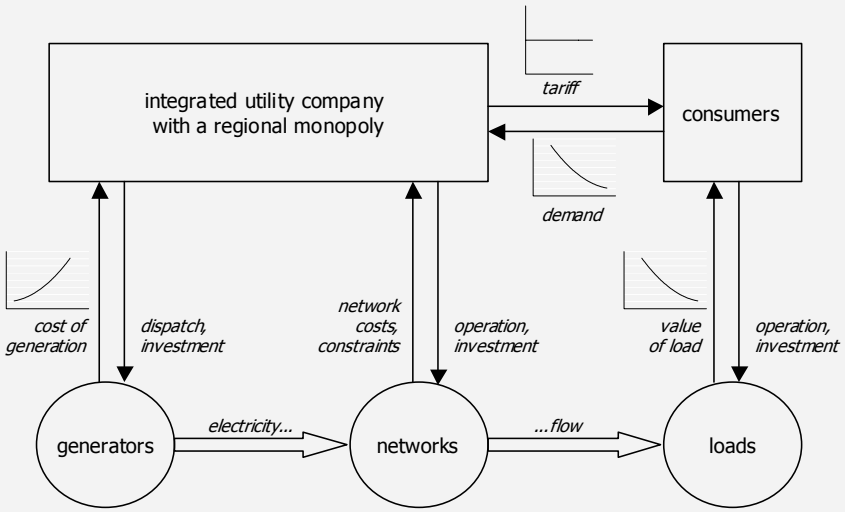


Figure 3.4: The electricity system before liberalization

3.4 The economic subsystem

3.4.1 Function and definition

Liberalization has broken up the centralized control of the utility companies. Whereas the economic 'layer' of the system used to be limited to the leadership of the integrated utilities (see the text box below), liberalization has created an elaborate structure on top of the physical system, which consists partly of a competitive market and supporting functions and partly of the network monopoly. Ancillary services for network operation can conceivably be traded in a market but they are often included in transmission services.

The economic subsystem is defined as the actors that are involved in the production, trade or consumption of electricity, in supporting activities or their regulation and their mutual relations. The economic subsystem controls the technical subsystem. The way in which it does this is the leitmotiv of this project. An essential feature of the economic subsystem is that it is constrained by the technical subsystem. Both subsystems are

constrained by regulations, such as construction permits, operating licenses and emissions permits for the technical subsystem, and competition law and EU directives for the economic subsystem.

It is important to make a clear distinction between the components of the two subsystems. The technical subsystem consists of components that either produce electricity, use electricity or are involved with the transmission of electricity. The instructions how to operate the system – the dispatch of generation, control of the transmission system – originate outside the technical subsystem. Before liberalization, a control center provided these instructions, both for generation (dispatch) and operation of the network. Liberalization delegated part of this function to market parties. Control of transmission and system operator remain centralized out of necessity but control has been shifted from the electric utilities to independent transmission system operators (one per control zone).

3.4.2 Actors

For any given physical electricity infrastructure, many economic structures are possible. In the last decade, several different market models have been tried, among others in England and in the Nordic countries. Undoubtedly, these models will evolve and new ones will develop. Depending on the model and its historic and legislative context, there may be different actors playing different roles in the market. Certain parties will always play a role, however: the main actors in the technical system also are important actors in the economic system. They perform the essential functions of supply, demand, transmission and distribution.

When a party is active in both subsystems, the roles it plays are quite different. In the case of a generating company, in the technical subsystem its role is to provide physical input to the system: to generate electricity. Therefore, only its generators are considered part of the technical subsystem. The company's other activities are not relevant to the technical subsystem. In the technical subsystem, activities are expressed in terms of quality and quantity of electricity generated, transported or used.

In the economic subsystem, on the other hand, the generating company acts as a supplier: it sells electricity for a certain price. How the company goes about producing electricity is not relevant to the market: only price, availability and reliability count. The only exception may be the existence of a separate market for green or sustainable energy. In the economic subsystem, the main variable is money. Technical characteristics only play a role to the extent that they restrict or dictate economic behavior (for instance in the case of a capacity shortage). The two subsystems are related but they are not linked one on one. A generator with a constant output may have fluctuating revenues as a result of variations in market price.

Similarly, the transmission and distribution networks are part of the technical subsystem but the network operators are players in the economic subsystem. Even though they offer a monopoly service, they are actors involved in the trade of electricity, as they influence the market through their network charges. Similarly, the system operator functions outside the market but is involved in the balancing of supply and demand and contracts

black-start capacity from the generating companies. In case of emergencies, the system operator may interrupt the power service to consumers.

In addition to the actors that play a role in both the technical and the economic subsystems, there are a number of parties who operate only in the economic subsystem:

- The market operator (MO) matches supply and demand by organizing markets, such as a spot market and term markets, and/or by coordinating energy programs by generators and loads.
- Traders buy and sell electricity in the various markets.
- Brokers arrange sales between various market parties.
- Retailers provide electricity to consumers, buying electricity wholesale, paying to use the transmission and distribution grids, and doing the billing.

In principle, the market parties determine the dispatch of electricity generation facilities. Contracts made in the (spot) market stipulate which generators will run at which times. Only if the network cannot safely accommodate the dispatch pattern may the transmission operator interfere. The system operator makes last-minute adjustments to maintain the energy balance.

3.4.3 Model

Figure 3.5 shows a model of the economic subsystem. The arrows indicate the direction in which electricity is sold. Producers may sell directly to large consumers, or wholesale trading companies may function as intermediaries. By definition, retail trading companies sell to small consumers. Wholesalers sell not only to large customers but also to retailers and other wholesalers. Market operators and brokers do not buy or sell electricity but perform a facilitating role. The system operator and the network operators, at the bottom of the model, provide transmission services and maintain the system balance. They are not market parties, as they have a natural monopoly, but they do perform an essential function in the market. In theory each market party could be allowed to buy transmission services but in practice access often is restricted to only producers and/or consumers.

Unbundling has separated the network activities from the other activities of the formerly integrated electric utility companies, such as trading. Because in these other activities the old utility companies do not differ functionally from other companies, they are not mentioned separately in Figure 3.5.

Even while Figure 3.5 is a simplified model, it indicates how easily the relationships in an electricity market can become complex. The electricity market may become intertwined with other markets, such as the gas market, further complicating relationships between market players. For the purpose of constructing an analytic framework, these links with other markets remain outside the scope of this chapter.

Two fundamentally different types of electricity markets exist. In markets with a mandatory pool, generators are required to offer all their electricity to a pool from which consumers (retail companies) obtain their electricity. The pool operator combines the functions of system operator and market operator: he takes care of both the economic

matching of supply and demand as well as the physical balancing of the system. Hunt (2002) calls this the 'integrated' market model. In this model, bilateral contracts between market parties essentially are financial arrangements that do not directly impact which generators run. The system operator dispatches generators to minimize overall cost, based upon the bid prices by the generating companies. This model is used in PJM, New York and New Zealand.

The alternative is what Hunt (2002) calls a 'decentralized' market in which the transactions between consumers and generating companies determine the dispatch of generation. In this model, the system operator only has the task of physically scheduling the market parties' contracts. His only involvement with the dispatch of generation is in the balancing market and, if necessary, congestion management. As the system operator does not operate a market (spot nor long-term), brokers and power exchanges may facilitate the matching of supply and demand. The analysis in this study is based upon the decentralized model, as it is the prevalent model in Europe.

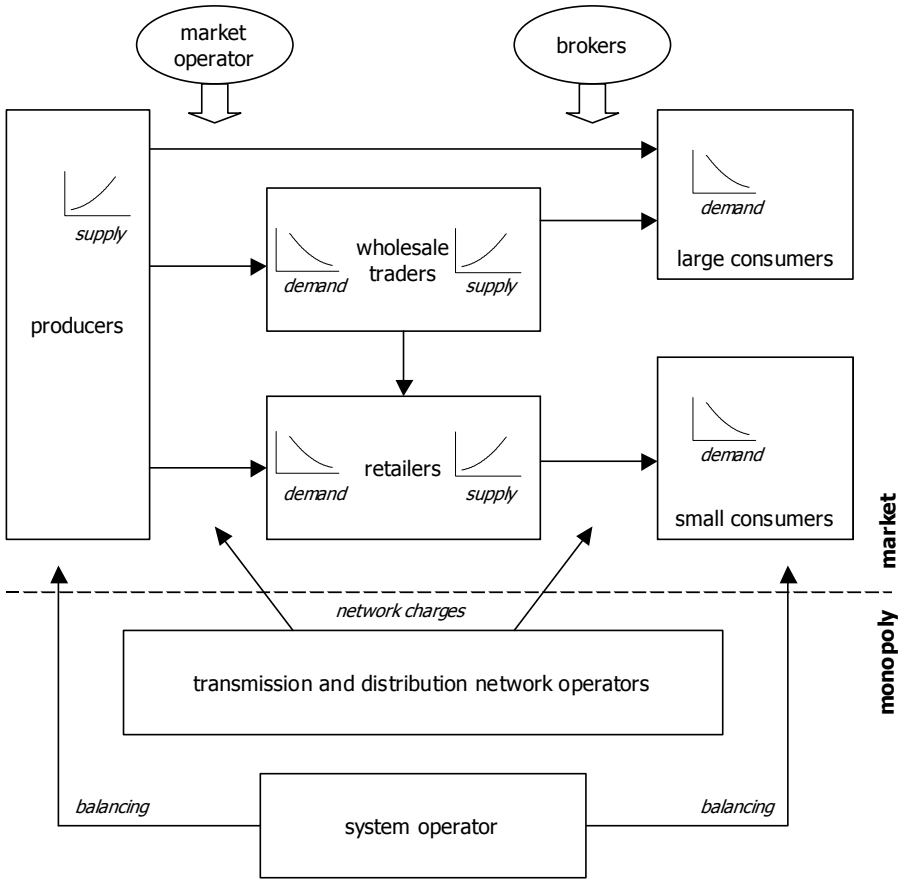


Figure 3.5: The economic subsystem

The representation of the market can be simplified to two groups, producers and consumers, meeting to match their supply and demand functions in the market, as shown in Figure 3.6. These groups are influenced by the system and network operators. In Figure 3.6, the arrows indicate information flows. In the 'matching' place producers and consumers may negotiate directly, with the help of a broker, in a spot market or through trading companies. How the matching process takes place is not important for the purpose of this chapter. What matters is that the market establishes a price for electricity and a related demand, and that the market decides how much of this demand is provided by which generation company. At least one of the parties involved in each series of transactions from producer to consumer must pay a transmission tariff to the network operator. The network operator may need to impose measures to manage congestion. The system operator balances physical supply and demand. Normally, he does this by adjusting generator output, to which end he contracts reserve capacity. In case of emergencies, if supply is not adequate to meet demand, the system operator may need to impose service interruptions.

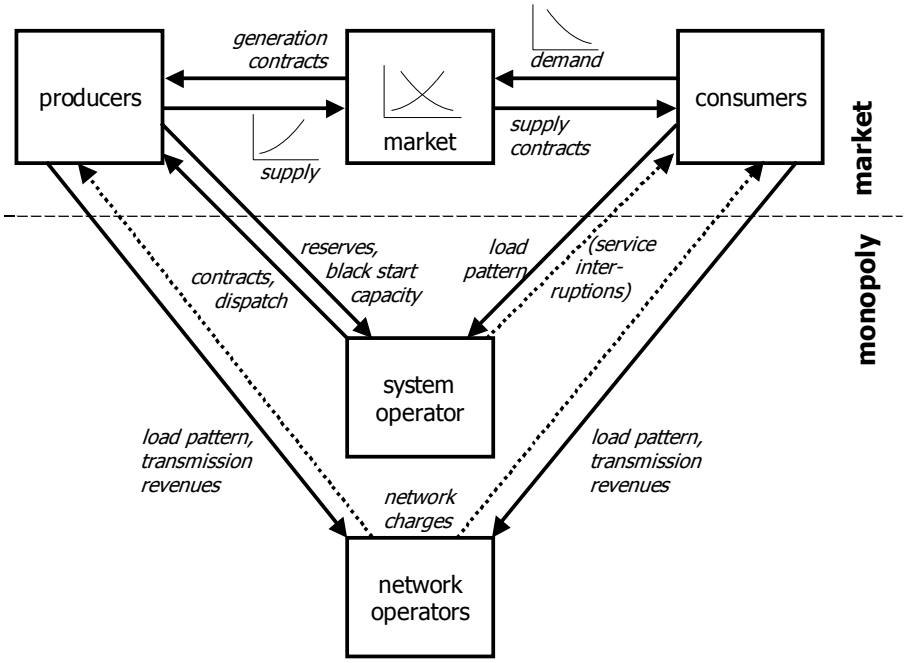


Figure 3.6: the economic subsystem, simplified

3.5 Links between the two subsystems

3.5.1 Links between the technical and the economic subsystem

The technical and the economic subsystems are linked by information flowing in both

directions in the form of prices, tariffs, willingness to pay, capacity restrictions and dispatch instructions, among others. In this section an inventory is made of the types of information that flow through the electricity system, beginning from the technical subsystem to the economic subsystem. Figure 3.7 presents a model of information flows from the physical infrastructure to the market. The transmission operator and the system operator have been merged into a transmission system operator. This simplifies the diagram and reflects the actual organization of European electricity systems.

The most important information for the market consists of the supply and demand functions. The supply function is based upon the cost function of generators but is not necessarily the same (indicated by arrow number 1 in Figure 3.7). There is a variety of reasons why the supply function is not necessarily the same as the cost of generation. Generators need to earn at least the marginal cost of generation to be active, but scarcity or the exercise of market power may lead to higher prices.

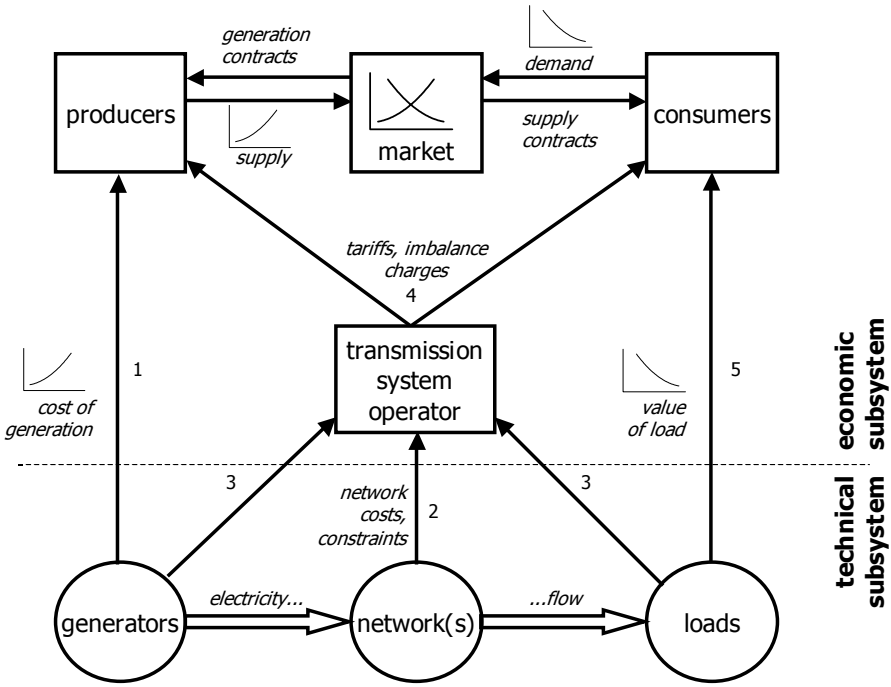


Figure 3.7: Information flows from the technical to the economic subsystem

The transmission system operator also charges market parties for his services but in a different way. The transmission system operator observes the network costs (arrow 2), which are the basis for the transmission tariffs (arrows 4), which typically reflect average cost. The tariff is regulated, publicly known and not negotiable, another difference with the price of generation. In addition to the price signals provided by transmission tariffs, the transmission operator may also signal capacity restrictions in the case of congestion. In his role as system operator, the TSO observes the balance between generation and load

(arrows 3) and arranges with generating companies and, if necessary, with consumers, to maintain or restore the balance. Consumers, finally, consider the value of their load and state their electricity demands to the market (arrow 5).

Summarizing, the market receives input in the form of supply and demand curves and transmission tariffs. Producers and consumers will negotiate, possibly through intermediaries, and a market price results. Based upon this price, consumers establish their demand for electricity. How much electricity is demanded, when and where and at which price, is the basis for the producers' decisions which generators to run.

3.5.2 Feedback to the technical subsystem

Now the feedback from the economic to the technical subsystem will be assessed. This feedback, for a large part in the form of operating contracts, is the dominating input for operational decisions in the short term and investment decisions in the long term, for both generation and transmission. Figure 3.8 shows a model of the possible links between the technical and the economic subsystems in both directions. The model in Figure 3.8 is an expansion of the one in Figure 3.7. The new arrows, printed in bold, indicate the feedback links from the economic to the technical subsystem.

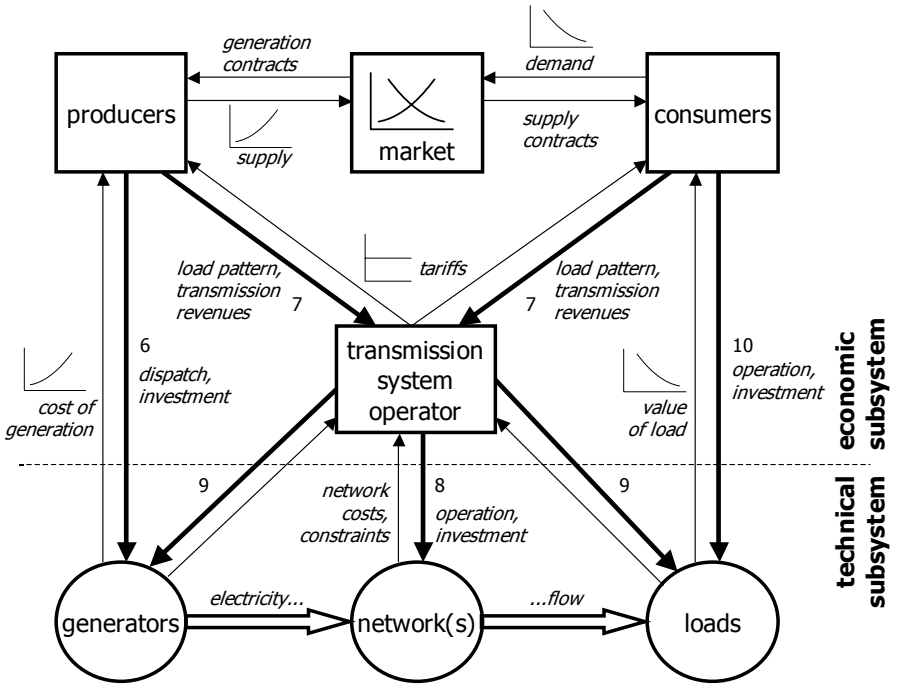


Figure 3.8: Relationships between the technical and the economic subsystems

The most important information that the electricity subsystem returns to the technical subsystem consists of the dispatch instructions (arrow 6 in Figure 3.8). Buyers of

electricity decide, using price and availability, from which producers they will purchase electricity. Price is to some extent influenced by transmission prices; availability is influenced by network constraints, in particular congestion. Electricity producers, in turn, decide which of their generators they will run to meet demand. Thus, the economic subsystem decides dispatch in two steps. Generators will also base their investment decisions on the prices they expect to receive in the market.

From the market parties' choice of producers and the producers' choice of generators follows the load pattern of the networks and the associated revenues to the transmission operators from the transmission tariffs (arrows numbered 7). This information is given to the transmission operators, who use the revenues to operate the network and invest in it (arrow 8). A special case arises when part of a network cannot accommodate the load pattern that follows from the transactions made by the market. This means congestion exists. The physical network cannot be expanded on short notice, so the transmission operator needs to use other means to solve congestion in the short run.

In the long term, the financial incentives that the transmission tariffs provide to the transmission operators are the only signal from market parties to the transmission system. It is a weak feedback loop, however, which means that there is a risk that transmission operators will not perform optimally. They are likely to over- or underinvest, depending on the specific incentive structure of the chosen model.

The transmission tariff system forms an essential part of the feedback loop to the entire technical subsystem. In addition to providing an incentive to the transmission operators, the transmission tariff system influences the electricity market by making certain generators more competitive than others. This dual effect of transmission tariffs is important. The level of transmission tariffs is a difficult tool for intervention in the electricity system by the government or the regulator because it influences two different parts of the system simultaneously.

The system operator may influence the generation and consumption of electricity (arrows 9) but his authority is restricted. He may pay or require generators to provide reserves and black-start capacity. With respect to consumers, the system operator's role is restricted to implementing service interruptions in emergencies and, in some cases, contracts for interruptible power as a form of system reserves.

The effect of the market upon electricity consumption appears smaller than upon generation. Consumers who do not have real-time meters, like many small consumers, typically do not exhibit much demand price-elasticity in the short run. Large consumers may show a more flexible demand curve, particularly if they are able to interrupt their production when prices rise too high. Both small and large consumers may show a long-term response by purchasing more energy-efficient equipment if prices are expected to be high (arrow 10).

Summarizing, the most important feedback from the electricity market to the technical subsystem is through dispatch instructions. By determining which generators will run, the load pattern of the network is determined. Dispatch decisions are taken in two steps: market parties choose generating companies who then select which of their generators

they will run. Dispatch decisions are based upon the combined prices of electricity and transmission and upon availability of generators and network connections.

3.5.3 Relevance of the model

While the model is not used explicitly in the first chapters, it was presented here because it represents the general way in which the electricity sector is regarded in this study. The issue of generation adequacy arises because the electricity sector is physically different from other markets. The main question here is which technical characteristics influence the market and how. Thus, the link between the technical characteristics of the system and the design of the market, which is the subject of the model, is central to the analysis. In Chapter 9, the coordination issue is described using the model more explicitly and developing it further. Here, not only the relationships between the technical and economic subsystems are at issue but also the relationships between the generation market and the networks.

3.6 System optimization

The question of system optimization is fundamental to this research project. The entire analysis is focused on the question of how the electricity system can be designed to best meet its policy goals. Therefore, some reflection upon what it means to optimize the electricity system with respect to these goals is necessary before the actual analysis is commenced. Making the concept of system optimization operational is not without difficulty. There are two basic aspects: what is the optimal system configuration at a given point in time and what is the optimal development path over time, given changes in demand, in environmental constraints, in the prices of primary energy sources, in technology, *et cetera*.

3.6.1 In theory

In the most general terms, economic efficiency means the maximization of the net social benefit (NSB) from electricity production, which is equal to the difference between the utility obtained from electricity U minus the cost of electricity supply C_{elec}:

$$Max\, NSB = Max[U - C_{elec}] \qquad (3.1)$$

The social cost of electricity supply as a function of output C_{elec} is the sum of the cost of the system, which can broadly be divided into the cost of generation C_{gen} and the cost of the networks C_{net}, plus external costs, such as environmental pollution C_{ext}.

$$C_{elec} = \sum C_{gen} + \sum C_{net} + \sum C_{ext} \qquad (3.2)$$

Ideally, net social benefit should be maximized over an entire interconnected electricity system and over time, as path dependency plays a significant role in the development of the system. The planning process that existed prior to liberalization included all the

factors in equation (3.2), at least in theory. In practice, insufficient incentives appeared to exist to minimize costs. Especially under cost-plus or rate-of-return regulation, the incentive to minimize capital costs is often too small.[2]

3.6.2 The use of constraints

There are several difficulties with finding the maximum for (3.1). First, the utility of electricity is difficult to establish. Even leaving aside the fundamental question of whether utility can be expressed properly in monetary terms, it is difficult to establish customers' willingness to pay for electricity. (See also Section 5.2.4.) In practice, the billing structure and the lack of information about real-time electricity prices cause consumer behavior to differ from what would be rational. As a result, it is quite difficult to establish what really is the utility provided by the electricity supply system, even when measured in monetary units. Consequently, the social cost of service interruptions cannot be established unambiguously, so it becomes impossible to provide electricity generators with efficient incentives for reducing outages to an efficient level with any degree of certainty. A solution is to simply establish a regulatory norm for reliability and use this as a boundary condition for the system.

The same argument holds with respect to the quality of the supplied electricity and other externalities: their social benefits, respectively costs, are difficult to establish. Again, a common solution is to apply standards, the effect of which is more certain. The industry may prefer standards, too, but for a different reason, namely if the alternative consists of Pigouvian taxation of externalities.

Consequently, the objective of maximizing net social benefit is often replaced in practice by a policy objective that is much easier to make operational, namely to provide electricity at the lowest possible cost within certain constraints. Then the welfare optimization function becomes

$$MIN\left\{\sum C_{gen} + \sum C_{net}\right\} \qquad (3.3)$$

subject to constraints concerning reliability, the quality of service, environmental standards, physical planning restrictions, safety standards, *et cetera*.

From the perspective of neo-classical economics, establishing regulatory standards is a step back from allowing the system itself to find an optimum because the probability is minimal that the regulatory standard (for reliability, for instance) is optimal. However, given the difficulties with establishing optimal financial incentives, for instance with respect to obtaining the necessary information, the standards may well be more efficient in practice.

Setting standards for reliability, the technical quality of the electricity and environmental externalities does not exclude the use of financial incentives for obtaining these

[2] Averch and Johnson (1962) published a seminal publication of this effect, which lead to an extensive volume of literature on the subject.

standards. It only means that, if incentives are used, they do not necessarily reflect real social costs. If incentives are used to reach, for instance, a certain standard of reliability, these incentives will need to be calibrated periodically. Just as the chosen level of reliability is not necessarily economically efficient, the incentives are not either. However, for the reasons mentioned above, this may still be preferable to trying to apply theoretically optimal incentives.

3.6.3 Unbundling and system optimization

Section 2.2.4 described why the unbundling of the network monopoly is a necessary condition for the introduction of competition in the electricity market. For system optimization, a consequence is that there is no opportunity to minimize costs in an integral manner anymore. Rather, competition puts pressure on generating companies to minimize the cost of generation, while pressure to minimize network costs must be created through some form of network regulation. Due to the physical close relations between generation and the networks, separate minimization of the costs of generation and those of the networks by no means automatically leads to overall system cost minimization. An important question for this research project is how sufficient coordination between the generation market and the networks can be maintained so total system cost is minimized and reliability targets are met.

Liberalization has transformed the objective of maximizing the net social benefit of the electricity system into the separate objectives of minimizing generation cost on the one hand and minimizing network costs on the other hand. The benefits of competition in generation come at the cost of having two sub-optimizations, one for generation and one for the networks, rather than a combined cost-minimization, plus a potential lack of coordination between the two. Whether there is an incentive for efficient coordination of generation and networks with the purpose of minimization of total system costs depends on the way the networks are regulated and the incentives that are given to generators through, for instance, the tariffs for the use of the networks. This issue is the subject of Chapter 9.

3.6.4 Dynamic optimization

The adequacy of financial incentives often is judged from a static perspective: in the absence of externalities, the market equilibrium is assumed to be socially optimal. This presumes, however, that an equilibrium will be reached. While in the short term the supply and demand of electricity usually are in equilibrium (or else the system would become physically unstable), a long term equilibrium between investment in generation capacity, network capacity and loads does not necessarily develop. The main obstacles are the long life-cycle of the hardware in the electricity industry, the path dependency in the design of the network and the often long lead times for installing new facilities such as power plants and transmission lines, while market conditions change continuously. As a result, even a system with a perfect incentive regime may lag behind the ever-changing demands that are placed upon it. As difficult as it may be to design a system that performs optimally according to certain fixed standards, the real challenge is how to achieve a long-term development path that minimizes the deviations from the social

optimum.

Optimality-based incentives

One approach is to try to establish an optimal set of incentives and constraints for the actors in the system and hope that this will minimize the system's deviations from the socially optimal state. The goal would be to design a system that tends to improve itself continuously as a result of each actor's attempts to improve his own situation. The presence of the correct incentives causes the actors in the system to make improvements continuously, even if an optimal state is never achieved. Therefore, the goal of achieving socially optimal investment in generation capacity is operationalized as structuring the system in such a fashion that each agent will tend to make socially optimal decisions given the current state of the system, existing information deficiencies and uncertainty about the future development of the system.

Asymmetric loss function

Not knowing what the future brings means that there is no analytical way to establish the optimal development path. Therefore it would be sheer luck if indeed the system developed optimally; in all probability, it will continuously deviate from the theoretical ideal. From this perspective, the policy goal for the electricity system should not be to achieve the social optimum, as it is unobtainable, but to minimize the social cost of erring. A relevant concept in this light is the asymmetric loss function (Morgan and Henrion, 1990). The social costs of deviating on one side of the optimum rise much faster than on the other side.

In the electricity system, this appears to be the case with respect to reliability (Cazalet et al., 1978). The social costs of providing too little generation capacity are much greater than the social costs of providing an equivalent level of generation capacity above the social optimum. This point will be elaborated upon in Section 5.4. When this is the case, the approach that minimizes the likely total loss of welfare over time may be one that consciously tries to overshoot the reliability target by a certain degree. This can be considered social insurance against the much graver consequences of a insufficient reliability of service. To implement this policy, the reliability standard need simply to be set higher than what is assumed to be optimal. Another approach is to include specific mechanisms in the design of the system that reduce the probability of under-spending on sensitive items, such as the capacity mechanisms in Chapter 6.

Summary

The two aspects of investment in generation capacity that are the subject of this study can both be regarded as optimization issues. This study takes as a starting point the approach to system optimization that is generally taken in practice, which is to operationalize system optimization as cost minimization within certain constraints. This approach is used for both the question of the optimal volume of generation capacity and the question as to how to coordinate investment in generation capacity with the development of the network. While there are numerous other constraints upon the system, such as environmental constraints, the focus is upon reliability. The reason is that the

organizational changes that are a result of liberalization may impact system reliability. Therefore, new methods need to be established for ensuring that reliability goals are met, whereas other, non-technical constraints such as environmental standards, can still be established in the same way as before.

A significant distinction exists between static and dynamic system optimization. Given the long lead times for new investments in the electricity sector and the long life cycles of components, the development of the system will always lag behind the changes in the demands which society places upon it. From this point of view, static optimization is a largely academic exercise, performed only because it is amenable to quantitative analysis. The main question, from the point of view of society, is how to minimize the social costs of the inevitable deviations from the optimum over time.

4 The electricity crisis in California

An analysis of the reliability of electricity markets, and especially one which focuses on the role of generation, cannot ignore the crisis in California's electricity market in 2000 and 2001. Not only is it a fascinating case study that provides interesting lessons, it also has become an international reference point for public policy. Policy makers of restructured electricity markets everywhere have asked how they can avoid such a crisis from happening in their systems. To answer this question, the root causes of the crisis need to be identified. This chapter concludes that the fundamental cause was a lack of investment in generation capacity. When the margin between supply and demand became slim, generating companies withheld generation capacity in order to push market prices above their competitive levels, severely aggravating the crisis. Consequently, the cost of the crisis to consumers was caused not only by the interruption of electricity service but also by the extended period of extremely high electricity prices.

4.1 Introduction

The electricity crisis that plagued California between the summers of 2000 and 2001 shocked and fascinated people around the world. How could such a high-tech state lose control of the electricity system to the extent that service could no longer be guaranteed? The disastrous developments in California caused widespread doubt about the desirability of creating markets in electricity. This chapter shows that California's problems can only partly be blamed on restructuring. On the other hand, some of the factors contributing to the crisis may also arise in other electricity systems.

The purpose of this chapter is to investigate the causes of California's energy crisis, not the consequences for the state or the electricity sector, nor the solution to the crisis. Other electricity systems are primarily interested in avoiding such a crisis, which the question how it developed is most interesting. An analysis of the crisis is particularly instructive for European electricity markets, which have in common with the pre-crisis market structure in California that they have no specific mechanism to encourage investment in

generation capacity. The solution to the crisis that was chosen in California is not evaluated, as the focus of this research is how to prevent such crises, not the question how to deal with them when they occur.

Section 4.2 describes California's electricity system in the years preceding the crisis, as well as the players and the rules of the newly restructured market. Section 4.3 presents a chronology of the crisis and reviews a number of relevant trends in the California electricity market. An analysis of the crisis follows in Section 4.4. The analysis is based in part on original data and in part on scientific literature and media sources. Section 4.5 summarizes the conclusions, while Section 4.6 draws some lessons for other electricity systems.

4.2 Restructuring California's electricity market

4.2.1 Prelude

Until 1996, electricity was provided in California by vertically integrated monopolies. The majority of the state was served by three investor-owned utilities, Pacific Gas & Electric, Southern California Edison and San Diego Gas & Electric. In addition, some cities provided electricity as a municipal service, the largest of which were Los Angeles and Sacramento. The privately owned utility companies were regulated by the California Public Utilities Commission, the municipal utilities were not. The utility companies, both private and municipal, produced much of their own electricity and owned the transmission and distribution networks. In 1978, the Federal Public Utilities Regulatory Policy Act partly opened the market for generation. In 1992, the Federal Energy Policy Act further opened the market in wholesale generation, among other things by facilitating access to the (mostly privately owned) transmission wires. The Energy Policy Act provided an impetus for the liberalization of the United States' electricity markets but it remained up to the individual states to take action. Nationally only about 7% of electricity was generated by non-utility companies by the year 2000 (Union of Concerned Scientists, 2000).

Prior to restructuring, the performance of California's electricity system was mediocre. Prices were high: by 1996, the average price in California was almost 40% higher than the average price in the rest of the USA (EIA, 2002). This was caused in part by large cost overruns in the nuclear power program (Hirst, 2001). Large businesses pushed for liberalization hoping that competition would lower the prices of electricity (Gladstone and Bailey, 2000). In the early 1990s, California was one of the first states within the USA to consider restructuring its electricity system. However, due to strong opposition, among others from Southern California Edison, the debate continued for a number of years before a compromise was reached in the form of Assembly Bill 1890 in 1996.

4.2.2 The rules

The California State Legislature adopted AB 1890 with the purpose of creating a competitive market in electricity. This act, which took effect in April 1998, applied only

to investor-owned utilities, which were under the jurisdiction of the California Public Utilities Commission. Municipal electricity utilities, the largest of which were those of Los Angeles and Sacramento, were not required to restructure and most of them did not. The investor-owned utilities served a little more than 80% of the California market (CEC, 2000).

In order to encourage the development of competition, the electricity industry was partially unbundled. The investor-owned utilities were pressed to divest their generation assets, other than nuclear and hydro power.[3] They now needed to purchase a large part of the electricity that they delivered to their customers on the market. For their captive consumers, they were required to purchase the electricity in a pool, the California Power Exchange (now defunct).[4] For most of its existence, the pool only allowed spot (day-ahead) contracts. Future and long-term contracts were not allowed in order to stimulate competition. However, an important side-effect of this rule was that it removed important risk management tools for market parties. The restriction of long-term contracts was changed in July 1999, when the power exchange started to offer a limited volume of long-term contracts.

While the utility companies maintained ownership of the transmission system, they had to hand over control of the network to the newly created Independent System Operator (ISO, also known as CAISO). Their role was reduced to distributing electricity, which they purchased from the pool or generated themselves, to their customers and managing their distribution networks. Hence, after restructuring the utility companies became known as utility distribution companies.

Rates for consumers were frozen at their 1996 levels and reduced by 10% in 1998. This rate freeze was instigated by the large utilities in order to recover their stranded investments (Gladstone and Bailey, 2000). The rate freeze was to remain in place for four years or until the utilities had recovered their stranded costs, whichever came earlier. The assumption was that competition would lead to decreasing wholesale prices, thereby increasing the utility distribution companies' profits. The difference between the fixed retail price and the wholesale price was the consumers' contribution to paying off the utilities' stranded investments, called the Competition Transition Charge.[5]

When AB 1890 entered into force in April 1998, most customers of the former investor-owned utilities were allowed to choose their provider of electricity. The law stipulated that the utility distribution companies had a service obligation only towards those customers who did not switch provider. In exchange for the opportunity of finding

[3] Note that in Europe unbundling is directed at separating networks, being considered a natural monopoly, from competitive functions, while in California unbundling involved the divestment of generation assets by the incumbent utility companies. The utility distribution companies were allowed to retain their networks as well as deliver electricity to the customers on those networks.

[4] The name California Power Exchange is confusing as the word exchange usually indicates that participation is voluntary while the term pool tends to be used for mandatory trade platforms.

[5] As wholesale prices started to exceed the fixed retail prices during the crisis, the CTC for the protected customers became negative. More simply said, the utility distribution companies lost money on their power sales, rather than recovering stranded costs.

cheaper providers, customers who switched to another provider did not enjoy the protection of fixed retail prices. As a result, few consumers switched so the utilities retained most of their customers and retail prices remained fixed for the great majority of consumers (Marshall and McAllister, 2000).

In summary, the main characteristics of the California restructuring process were:
- the utilities divested much of their generation assets;
- transmission was managed by an Independent System Operator;
- most purchases by the utility distribution companies were made from a mandatory power pool in which only spot contracts were allowed;
- during the transition period (which lasted through the power crisis), retail prices were fixed for small consumers;
- not all of California participated in the restructuring process: a number of municipal electricity departments stayed out.

4.2.3 The players

The restructured part of the California electricity market was served by three large utilities. Restructuring changed these formerly vertically integrated utilities into utility distribution companies:
- Pacific Gas and Electric served about 12 million people in the north and middle of California with gas and electricity. After restructuring, it owned about 7500 MW in generation capacity in the form of hydropower and a nuclear plant. It had to purchase about 60% of its 82000 GWh of annual electricity sales on the wholesale market (Pacific Gas and Electric, 2001).
- Southern California Edison was nearly as big as Pacific Gas and Electric. However, it owned less than 3000 MW in generation facilities, consisting mostly of a majority stake in a nuclear power plant.
- San Diego Gas and Electric sold only a little less electricity than its two counterparts but owned almost no generating facilities (CEC, 2001a).

The generation assets that the utility companies had sold were purchased by private companies. After restructuring, about 40% of the electricity sold in the state was generated by private firms, the largest ones of which were AES, Reliant, Southern (Mirant) and Duke. Public agencies such as municipal utility companies still produced nearly a quarter of all electricity. A slightly smaller percentage was provided by 'qualified facilities,' which were Federally approved environment-friendlier facilities such as combined heat-and-power units and renewable energy plants. The utilities were required to buy from these qualified facilities, which typically were small and independently owned. Finally, the utilities themselves provided about 15% of total production (Kahn and Lynch, 2000).

The California Public Utilities Commission regulated the tariffs that were charged to captive consumers, not only for electricity but also for other privately owned utilities such as gas, water, and rail transport. After restructuring, its role in the electricity market was limited to setting the rates for small consumers as long as they remained captive.

The California Energy Commission (CEC) was an energy planning agency that survived

the restructuring process. Its main activities were data collection and analysis and power plant licensing. It provided much of the data for this case study.

The restructuring law created two non-profit agencies to facilitate the operation of the market. The Independent System Operator took over control of the transmission network from the incumbent utilities. The California Power Exchange became the trading platform for most wholesale trade. The distribution companies were required to purchase much of their electricity there, so it actually was a mandatory power pool for most transactions. These two organizations were overseen by the Electricity Oversight Board, as part of its more general monitoring function.

Restructuring not only involved a shift of control from the State to the market but also to the Federal government. The State had surrendered its authority to intervene directly in the market to the Federal Energy Regulatory Commission (FERC). The State's only power to intervene in market operations lay in the Electricity Oversight Board's ability to litigate with the FERC. However, during the crisis the FERC adopted a laissez-faire approach until political pressure forced it to change its course.

4.3 Crisis

4.3.1 Chronology

The electricity system functioned smoothly during the first two years after restructuring took effect in April, 1998. Wholesale prices dropped and remained fairly stable so the restructured utility companies made a profit. In addition, they received substantial revenues from the sale of their fossil fuel generation assets. During the crisis, when the utilities asked the state for debt relief, these revenues became a point of contention as they had been transferred to the utilities' parent companies and were not used to offset the utilities' mounting losses.

In May, 2000, two years after the restructuring act had taken effect, wholesale prices started to rise sharply, as Figure 4.1 shows. Due to the demand for air conditioning, in most of California electricity demand peaks in the summer. A certain price increase was therefore to be expected but during July and August, the average electricity price at the power exchange was more than three times higher than at the same time in previous years.

San Diego Gas & Electric was the first of the three investor-owned utilities to recoup its sunk investments and therefore was the first utility whose retail rates were freed. In the summer of 2000, shortly after their retail prices were deregulated, SDG&E consumers saw the amount of their power bills multiply. This caused a political uproar that led the State Public Utilities Commission to freeze the rates of San Diego Gas & Electric again. The combination of fixed retail rates and high wholesale prices caused all three utility distribution companies to lose money but the expectation was that prices would decrease after the summer and things would return to normal again.

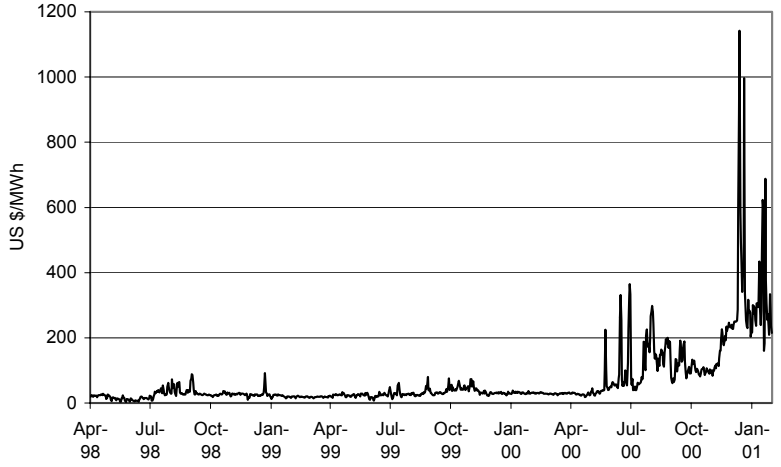


Figure 4.1: Daily average wholesale prices
Based upon data from the University of California Energy Institute

Many causes were suggested for the price increases but it was clear that there was an acute shortage of power on the market. This shortage reached its first climax with rolling black-outs in the San Francisco Bay area in June, 2000. This outage was not actually caused by a general shortage of generation capacity but by a local shortage combined with a lack of sufficient transmission capacity (Johnson and Woolfolk, 2000).

To much surprise and dismay, prices did not return to normal after the summer peak but remained much higher than usual during the fall of 2000. The shortage continued even though demand declined. The continuing high prices caused the utilities' financial losses to grow quickly as their obligation to serve their customers at fixed tariffs remained. The utilities agitated for a lifting of the fixed retail tariff, or at least for a rate increase, but as elected bodies, the legislature and the California Public Utilities Commission were extremely reluctant to do so. They continued to expect prices to drop, so the utilities could recover their losses. To reduce the gap between the utilities' purchase prices and retail prices, California appealed to the FERC to impose regional wholesale price caps but the FERC refused (Allen and Booth, 2001). State-imposed price caps were ineffective due to generators' ability to 'launder' electricity by selling it to affiliates in neighboring states and buying it back at much higher prices. As wholesale prices remained high, the combined losses to the two largest utilities exceeded USD 12 billion by February 2001, bringing them to the brink of bankruptcy.

The utilities' insolvency compounded the problems to the point that at the end of 2000 the situation became untenable. The large private generators were threatening to stop delivering electricity to the utilities because the latter's creditworthiness had sunk to the lowest level. In addition, the utility distribution companies had not been able to pay the small, independent producers for so long that some of them could not afford to purchase fuel any longer and risked bankruptcy themselves. The utilities' lack of creditworthiness

also threatened their ability to purchase natural gas and deliver it to generating companies. These effects further jeopardized the supply of electricity. In December, the outgoing Clinton administration intervened by ordering generators to keep supplying electricity whether they were paid or not. After President Bush took office in January, he continued this order for a few weeks but gave California notice that it must devise its own solution.

In January 2001, the crisis reached a new climax when the independent system operator was forced to impose rolling blackouts throughout significant portions of the state for two consecutive days. At the same time, prices at the power exchange rose several times higher than the previous record highs, causing the collapse of the utility distribution companies. Towards the end of January, the State of California finally took decisive action. Effectively abolishing the market altogether, the state started purchasing electricity on behalf of the teetering utilities. As the state became the single buyer of electricity, the power exchange, the credibility of which already had been undermined by allegations of manipulation, closed its doors.

The state had hoped to be able to use its purchasing power to drive prices down but it paid two to three times the historical average electricity price for long-term contracts. In this way it not only guaranteed the flow of electricity but also ensured that Californians would be paying for the crisis for years to come (Nissenbaum et al., 2001). The cost of buying power was high; it was estimated that the cost to the state was USD 18 billion in the first year alone, which caused Wall Street to reduce the state's credit rating (Nissenbaum, 2001). The state intervention stabilized the supply of electricity but at a high cost.

There were several more black-outs in the period through May but by the summer of 2001 the crisis ended. For a large part, this was probably due to Californians' energy conservation efforts, which reduced peak demand by up to 12%. In addition, mild weather played a role and all possible means were developed to increase production. By the summer of 2001, more than 6000 MW in new generating plants had already been approved for construction while the same amount was under review (CEC, 2001b).

During the crisis, a heated debate developed with respect to its causes and potential remedies. Opponents of deregulation were quick to claim that the crisis was proof that a market in electricity was impossible. Proponents of deregulation countered that the way in which California had restructured was highly flawed so it could not be considered proper deregulation. To begin with, they considered the fixed retail prices and the prohibition of long-term contracts at odds with a competitive market (Manifesto, 2001). Opponents of environmental measures claimed that California's relatively strict environmental regulations had made all investment in power plants impossible. The state accused generators of price gouging. The generators washed their hands in innocence while making record profits, attributing the high prices to market conditions (Hebert, 2001; Holson and Oppel, 2001). Other suggested causes were high gas prices, high prices of NO_x emissions permits, low reservoir levels due to drought, falling imports from other states and an unprecedented growth in demand due to the fast development of the computer and internet industry.

4.3.2 Trends

To evaluate the different possible causes of the energy crisis, it is useful to separate physical reality from the many theories and accusations that were suggested during and following the crisis. This section analyzes the physical factors that determine reliability, in order to help explain the development of the crisis: demand, installed generator capacity, generator outages and imports.

First, the claim that the state was confronted with an extremely high and unforeseeable growth in demand appears to be a myth. As can be seen in Figure 4.2, electricity demand grew at a modest annual 1.4% on average during the decade preceding the crisis, trailing population growth slightly. This means that per-capita electricity consumption decreased marginally despite the economic upturn. The growth in annual peak demand was also not substantial, according to the CEC (2000). In the late 1990s, peak demand in the ISO control area grew to about 46000 MW. While there were regional variations within the state, the overall modest growth rate indicates that demand-side developments were not the cause of the general power shortage.

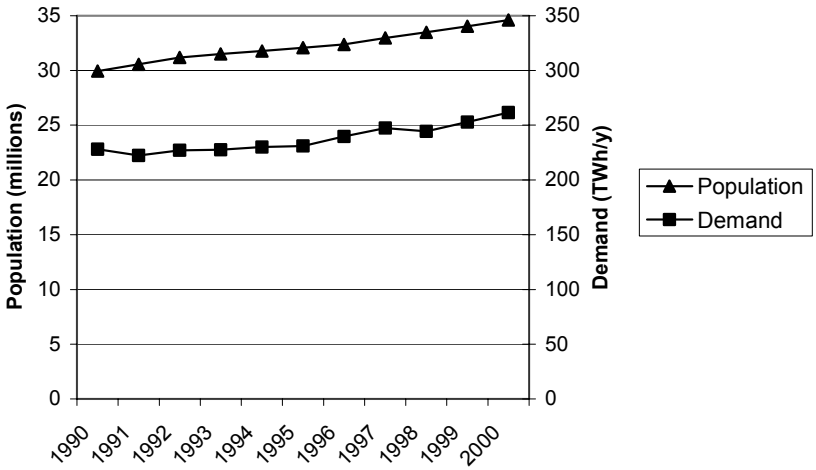


Figure 4.2: Electricity demand and population growth
Source: CEC, 2001a

On the supply side, a remarkable trend was that very little generation capacity was added during the 1990s. As a result of decommissioning, the net available capacity actually decreased slightly during the decade prior to the crisis (World Bank, 2001). In the summer of 2000, the installed generation capacity was close to 53,000 MW (CEC, 2001a). The actual available capacity in California was substantially less due to scheduled and unscheduled outages. When comparing generation capacity to demand, it should be kept in mind that an electricity system needs reserve capacity to function. In California, operating reserves of 7% are the norm for thermal capacity and 5% for hydro capacity (California ISO, 2001).

The growth rate of generation capacity lagged substantially behind the growth in electricity consumption not only in California but also in the rest of the Western System of the USA. Throughout the 1990s, the demand for electricity in the West grew on average by about 3% per year while generation capacity in that area increased by less than 1% per year (FERC, 2000). As a result, the margins between available generation capacity and demand shrank throughout the Western System (Weare, 2003).

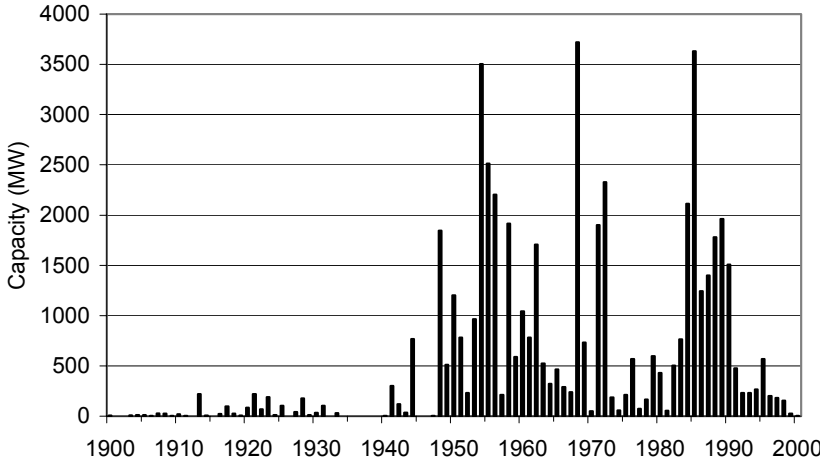


Figure 4.3: Generator construction year
Source: CEC 2001a

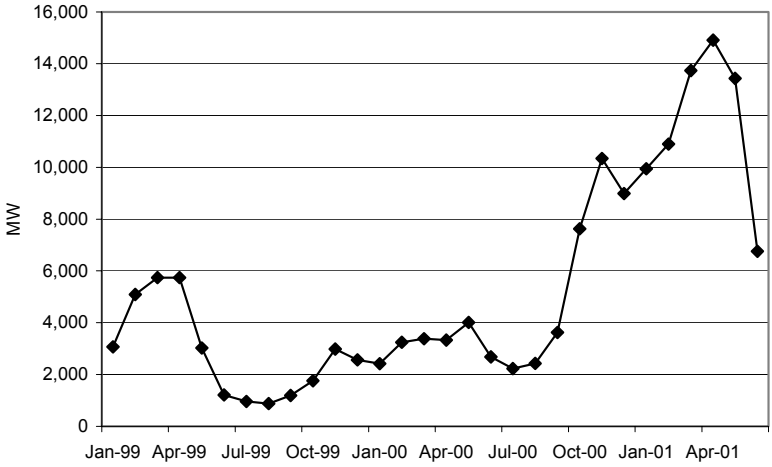


Figure 4.4: Monthly average off-line capacity
Based upon data from the California Energy Commission
(www.energy.ca.gov/electricity/1999-2001_monthly_off_line.html)

The scarcity may have been made worse by the high average age of the available generating stock which probably reduced plant reliability. Figure 4.3 shows that nearly all existing plants were built before 1990 and that about a third was from before 1970. Historically, the ISO could assume that about 2500 MW was out of service at any time. During the crisis, the volume of off-line capacity grew to unprecedented rates, as is shown in Figure 4.4. While the outage rate was somewhat low during the summer of 2000, during much of the winter more than 10,000 MW was off-line and at one point the outage rate exceeded 30% of total installed generation capacity.

The last factor to be considered is California's reliance upon electricity imports. Figure 4.5 shows that California relied substantially upon imports to meet its electricity demand during the last two decades. While California's relative dependence upon imports has been declining (as the absolute amount remained stable), the state was still highly dependent upon imports at the time of the crisis. In 1999, the share of imports in total consumption was 18%. Imports typically are higher during the summer, when the Pacific Northwest has its off-peak season and can export more of its large supply of hydropower.

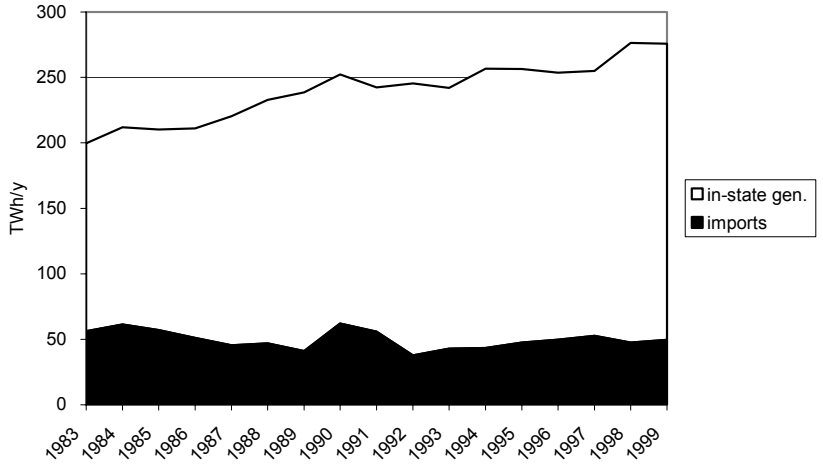


Figure 4.5: Imports and in-state generation
Based upon data from the California Energy Commission

In 2000, net imports were 28% less than the year before (California ISO, 2001). There are two main causes for these import reductions. In 2000, a drought hit the northwest as a result of which hydropower production was down significantly and the area could export much less than usual. In the winter of 2000-2001, the Bonneville Power Administration had about 4000 MW less capacity than usual due to the drought, which was one of the worst in 70 years of record keeping (Bonneville Power Administration, 2001). In addition, strong economic growth and population growth in surrounding states coupled with lagging investment, caused these states' power surpluses to shrink.

In summary, the image that emerges from these trends is that the capacity margins (the

difference between generation capacity and imports on the one hand and demand for electricity on the other hand) steadily shrank throughout the 1990s. Demand did not grow exceptionally fast (neither peak nor average demand) but it did grow. This growth was not matched on the supply side where both generation capacity and imports remained stable. During the crisis imports dropped significantly and generator outages increased dramatically.

4.4 Analysis

4.4.1 Physical crisis

The question that the California crisis provoked in every other restructured electricity market was 'Can it happen here?' To find the answer, it is essential to understand what went wrong in California. In particular, the question should be addressed which aspects of the crisis were unique to California or simply the result of bad luck, and which factors may occur in other systems as well.

A basic flaw in the California market was the fact that prices were fixed for a majority of consumers. Not only did this cause the financial downfall of the utility distribution companies, it also meant that consumers did not react to the high electricity prices.[6] Had consumers been able to react to the electricity price, blackouts might not have occurred nor would prices have risen so high. However, the lack of consumer responsiveness is not unique to California. In other systems, in which consumer prices are free, a majority of consumers rarely know the real-time price of electricity and have no incentive to react to it. A common market flaw is that the time of electricity consumption is not metered. As a consequence, the price paid by consumers is averaged over the 24 hours of the day and among many consumers as it is not known which consumers contribute most to the expensive demand peaks. Had the consumers in California been exposed to real market prices, they would still have received an incentive to reduce their electricity consumption only when average prices rose. There would have been no incentive to reduce consumption when it was needed most, during peak hours. The crisis in California probably would have been alleviated substantially if consumers had shifted consumption from peak hours to off-peak hours but in the absence of time-of-use billing there was no reward for doing so.

It cannot be assumed that the problem of the low demand price-elasticity was the only cause of the crisis in California. Even with full demand-side participation, generation capacity still needs to be adequate to meet demand peaks. We need to turn to the supply side of the market to investigate why it failed to produce enough generation capacity to meet demand. California started from a situation of plenty reserve capacity in the late 1980s but during a decade with no net investment in generation the reserve capacity was absorbed by the growth in demand. Even when demand grows as moderately as it did in

[6] Only after the rolling blackouts occurred and a public outreach campaign was started, supported by some financial incentives for conservation, did consumers start to reduce their demand. This was, however, not a reaction to the actual prices in the market.

California, a stagnation of investment in supply capacity is bound to cause a shortage sooner or later. This aspect of the crisis will be called the *physical* crisis as opposed to the *financial* aspect of the crisis, a distinction that The Brattle Group also makes (Carere et al., 2001). It is often suggested that environmental restrictions and NIMBY-type opposition made it difficult to obtain building permits for power plants in California (cf. Yardley, 2001; Berry, 2001). However, the fact that some projects for which the necessary permits had been acquired were not built, suggests that other reasons were dominant (Carere et al., 2001). Why, then, was there no new construction until it was too late?

One apparent reason for the lack of power plant construction was regulatory uncertainty (Carere et al., 2001; Hirst, 2001). The debate about restructuring started in the early 1990s while the law took force in 1998, so during most of the decade the future of California's electricity system was unclear. Since restructuring, investors may have been reluctant to invest in new plants until they had developed some experience with the new system. The forced absence of long-term and future contracts eliminated possibilities to hedge the investment risk in generation capacity, which further discouraged new construction.

Low prices before the crisis discouraged the development of new generation capacity. Due to the lack of experience with the market, investors did not know when to expect price rises nor did they probably realize how high the prices could rise during a shortage. Once the prices rose, there was not enough time to construct new capacity. Planning and building a new generation plant may take several years; the permit process may extend the lead time with again as much time. Investment in generation capacity in reaction to price peaks is slow therefore, as a result of which there is a risk of cyclical investment behavior. The possibility of the development of a business cycle was anticipated by Ford, based upon a simulation of the California electricity market (Ford, 1999; Ford, 2001). The theory that electricity markets are susceptible to a business cycle was further developed by Stoft (2002).

The fact that there was little investment in generation capacity throughout the western states means that the cause of the crisis was not limited to California alone. A system-wide lack of investment, coupled with often higher growth rates of electricity consumption in surrounding states, actually caused a regional crisis (Hirst, 2001). The reason that the crisis manifested itself primarily in California is that California relied heavily upon imports. When surrounding states saw their electricity surplus disappear, they reduced their exports. Why the surrounding states did not invest was not investigated in this project. Perhaps regulatory uncertainty also played a role there, as some states were also (considering) restructuring. On the other hand, they simply may not have been interested in constructing generation capacity as long as they had more than enough capacity to meet their own demand. The NIMBY phenomenon may have played a role, as it is difficult to make the case for power plant construction in a state with a power surplus.

There were many factors that exacerbated the physical crisis, such as the drought, which reduced the supply of hydropower (much of which was imported), the hot summer of

2000, which increased consumption and technical problems with California's many old generators. However, these same circumstances would not have caused shortages a number of years earlier when the system still had enough reserve capacity. The lack of investment in generation capacity brought the system to a point where it could no longer cope with irregular circumstances. The reason the crisis occurred in the summer of 2000 is that a number of unusual circumstances occurred simultaneously. Had the summer of 2000 been wet and cold, the crisis might have been delayed until a year later. However, a shortage of electricity was bound to occur sooner or later, as long as no new generation capacity was being built.

An aggravating circumstance that stands out is the shortage of transmission capacity. The San Francisco Bay area blackouts in June, 2000 could have been avoided if the transmission capacity into the area had been larger. At the time of these first blackouts, the state actually had a power surplus but the necessary power could not be transmitted to the Bay area. The problem with transmission was not just one of capacity: due to the deteriorating quality of the networks, the net available capacity had been falling gradually since the late 1980s (Hirst, 2001).

Given the lack of generation capacity, physical shortages would probably also have occurred even if the electricity industry had not been restructured. Generating companies might still have underestimated the need for new capacity until it was too late, with power interruptions as a result. The financial chaos that ensued, however, was a product of the new organization of the sector.

4.4.2 Financial crisis

In any market, scarcity can be expected to increase prices. The situation in California was exceptional, however, in two respects: wholesale prices rose to extreme heights for a long period of time and retail prices remained fixed. This particular combination caused a financial crisis, the damage of which may rival the costs of the actual power outages. Since the beginning of the summer of 2000, wholesale prices exceeded retail tariffs by a large margin. At times, wholesale prices rose to more than ten times the retail price. The utility distribution companies, who had few long-term contracts and an obligation to serve their small customers, were forced to purchase their electricity at these high prices and sell it at the fixed, low retail price. As a result, the utilities' losses quickly mounted to billions of dollars. By early 2001 they became insolvent, causing a collapse of the system.

The utility distribution companies' deteriorating financial situation complicated and aggravated the crisis. The utilities' suppliers became reluctant to deliver electricity and gas to them, which further threatened the delivery of energy to their customers in California. In addition, the utilities' months-long failure to pay small, independent producers for the electricity they delivered drove these producers to the brink of bankruptcy, posing another threat to the supply of electricity. Thus, the financial crisis contributed to the physical shortage of electricity, creating a feed-back loop that worsened the crisis.

The financial crisis contributed substantially to the social costs of the crisis because it left California without a functioning electricity market. This forced the State of California to intervene by purchasing electricity on behalf of the utility distribution companies. Not having any experience in the market and entering it when prices were high, the state engaged in many long-term contracts at prices far above the average market price.

The financial crisis raises two questions: why were the retail rates fixed and why were wholesale prices so high? The first question is the easiest to answer. The utilities themselves had lobbied for fixed retail prices, ironically because they thought this would protect their profit margin. Assuming that the market would bring wholesale prices down, they hoped to avoid a price war and protect their profits by fixing retail rates during a transitional period. The legislature's justification for permitting this was to allow the utilities to recover their stranded investments.

The second question is more complex. When a shortage of supply exists in a market, one can expect prices to increase, especially in an electricity market where demand is characterized by low price-elasticity. However, it is questionable whether this effect alone can explain prices of ten times the historical price. A first explanation is that a number of input costs had increased. The price of natural gas in California soared in the fall of 2000 (CBO, 2001). At first, this appeared to be caused by scarcity due to the high demand from electricity generators and a break in a pipe line. Later, however, evidence emerged that market manipulation by gas supply companies caused at least part of the increase in gas prices (Sheffrin, 2002; Oppel and Berman, 2002; FERC, 2002a). Other input costs rose as well. The reduced availability of hydropower due to the drought meant that more electricity had to be generated from fossil fuel plants, as a result of which the price of NO_x emissions increased substantially (Joskow and Kahn, 2002).

Bushnell (2004) argues that the near-complete absence of forward contracts was the main cause of the crisis. Forward contracts would have provided a double benefit: they would have provided the distribution companies with a hedge against high prices in the power exchange, and they would have reduced the incentive to generating companies for withholding capacity. (See the next section.) Part of the reason that there were so few forward contracts is that they were actively discouraged by the California Public Utilities Commission. Bushnell (2004) argues that a lack of foresight and insufficient incentives to the distribution companies for engaging in long-term contracts also played a role. The companies may not have believed that power exchange prices could rise as high as they did, or they may have believed that if it happened, the retail rates would be adjusted accordingly.

An extensive analysis of input costs and generator availability to determine the causes of the high electricity prices in the summer of 2000 shows that the high electricity prices cannot fully be explained by the cost of production and the inelasticity of the demand curve (Joskow and Kahn, 2002). The unusually high volumes of unavailable capacity during the crisis (see Figure 4.4) gave rise to the suspicion that generating companies were withholding capacity in order to increase the prices.

4.4.3 Manipulation

The suspicion of capacity withholding arose even as the crisis unfolded. It was counter-intuitive that the first state-wide electricity shortage would develop in January, during the low season. Generating companies argued that the high outage rates at that time were caused by deferred maintenance during the previous half year, when the continuing shortages had required the utmost from the aging generation facilities (Kaplan and Guido, 2001). Evidence of capacity withholding mounted, however. Joskow and Kahn observe that, in a situation of scarcity and with low price-elasticity of demand, generators have market power, that they are in a position to abuse this market power even without needing to collude formally, and that they have strong incentives for doing so. Using publicly available data, they show that capacity withholding is the only plausible explanation for a significant part of the high prices. Therefore, they conclude that the abuse of market power contributed significantly to the crisis. Stoft corroborates this analysis in a more theoretical way. He shows how a generator with even a small market share is able to increase its profit by withholding part of its generation capacity during a period of tight supply (Stoft, 2002). More evidence of the abuse of market power emerged from the Enron bankruptcy proceedings, where memos were found that explained the different methods that were used to drive up the prices (Behr, 2002; FERC, 2002a). The California Public Utilities Commission concluded eventually that a majority of the outages was caused by strategic withholding with involvement of all major independent electricity generation companies (CPUC, 2002). Recently, indications emerged that the prices for the NO_x emissions credits were also manipulated in order to manipulate electricity prices (Kolstad and Wolak, 2003).

Figure 4.6 shows electricity prices in California in June of 2000 in relation to the available generation capacity. Price is on the vertical axis, generation capacity on the horizontal axis. The vertical, dotted line indicates a situation in which demand equals available supply; the area to the right of this line indicates absolute supply shortages. (These do not necessarily lead to supply interruptions, as the system operator may be able to use his operating reserves to maintain the system balance.) In a competitive market, one would expect prices to rise as the generation capacity margin shrinks. However, the figure shows that high prices also occurred when there was ample generation capacity, suggesting the presence of market power among electricity producers.

The withholding of generation capacity by the electricity generating firms severely aggravated both the physical and the financial aspects of the crisis. It greatly increased the social cost of the crisis and contributed to the bankruptcy of the utility distribution companies. It would be a mistake, however, to consider the manipulation the main cause of the crisis. In the absence of timely investment in generation capacity, a shortage of electricity would have occurred sooner or later, if the market had not been restructured. And given a scarcity of electricity, wholesale prices would have risen and the fixed retail prices would have caused heavy losses for the utility distribution companies even if the market were perfectly competitive. Had the retail prices not been fixed, the utility distribution companies would have been able to pass the high wholesale prices along to consumers and avoid losses. However, evidence from San Diego suggests that unregulated retail prices might instead have caused a political crisis of at least the same

proportion.

Nevertheless, the risk of generation capacity withholding is an important lesson for other electricity systems. This phenomenon is by no means constrained to California but can occur in every electricity system with a low price-elasticity of demand when generation capacity is scarce. The experience in California shows that capacity withholding can inflict high social costs because it not only leads to excessively high electricity prices but may also contribute directly to service interruptions.

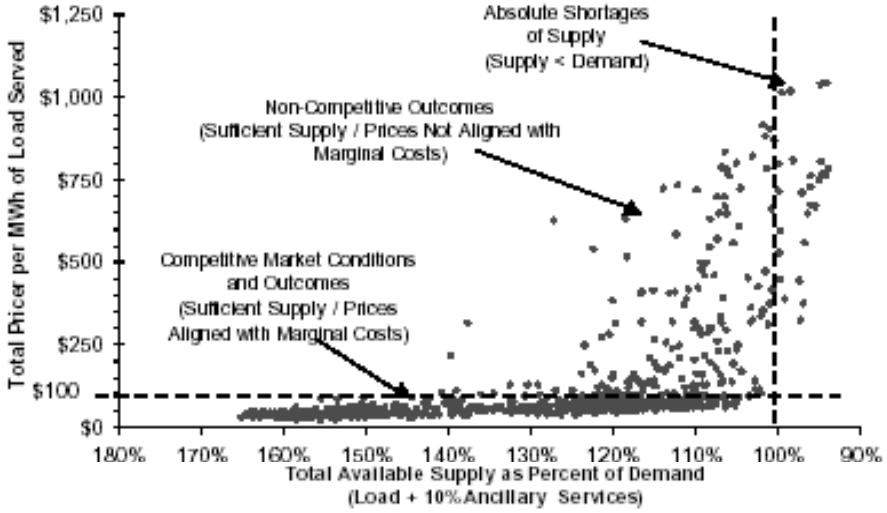


Figure 4.6: Market prices versus supply adequacy in California in June 2000: signs of price manipulation
Source: California ISO, 2000

4.5 Conclusions

The fundamental problem of California's electricity market was a lack of investment in generation capacity. For too long, too little was invested in generation capacity, not just in California but also in the other states of the Western interconnected system. The shortage of generation capacity was aggravated by a number of circumstances such as drought (which reduced the available hydropower), heat (which temporarily increased demand), generator failures and transmission capacity constraints. This inevitably led to a crisis in which the supply of electricity was not sufficient to meet demand. As the largest net importer of electricity in the region, California bore the brunt of this crisis when drought reduced electricity generation in neighboring states.

The pattern of underinvestment predated the restructuring law, which means that the crisis cannot be blamed entirely upon the market reforms. However, the new market structure did fail to signal the need for new generation capacity in time. The most likely

cause of the lack of investment within California is regulatory uncertainty during the time leading up to the restructuring of the market, exacerbated by a flawed market design. In addition, generating companies may have reduced their investment risk in the newly liberalized California market by waiting until the demand for new capacity was manifest. However, the long lead time of new generation capacity meant that investments in reaction to the shortage arrived too late. A consequence of the long lead time for new generating facilities plus volatile market prices is that a boom and bust cycle may develop.

A fundamental flaw in the restructuring law was that retail prices were fixed while wholesale prices were determined by the market. Combined with a lack of forward contracts, this caused the physical shortage to develop into a financial crisis. California's electricity distribution companies were forced to purchase electricity on a wholesale market in which, during the crisis, prices rose much higher than the regulated retail tariffs for which they had to sell much of their electricity. The fact that utilities had not been able to hedge their price risk by engaging in long-term power contracts aggravated the situation. The combination of high wholesale prices and fixed, low retail prices resulted in high financial losses for the utility distribution companies which eventually brought them into severe financial difficulties, causing the collapse of the market.

Prices rose to extreme heights during the crisis for a number of reasons. First, electricity demand in California was nearly price-inelastic, as a large portion of consumers purchased electricity for fixed tariffs. Thus, high prices did not lead to a reduction in demand, which could have dampened the price increases. Second, the costs of certain inputs, most notably natural gas and NO_x emissions credits rose substantially during the crisis. (Both appear to have been affected by the abuse of market power.) Finally, the abuse of market power severely aggravated the crisis. By withholding generation capacity, generating companies increased prices to far above their competitive levels and contributed to the power outages.

4.6 Lessons for other electricity systems

Several lessons can be learned from the experience in California. First, it is essential for the long-term stability of the system to have adequate incentives for investment. In California, these incentives existed – witness the high prices – but did not manifest themselves in time: only when there already was a shortage did the market take the initiative to add generation capacity. The consequence was that there was a period of tight supply before new capacity became available. This raised the possibility of a business cycle.

The liberalization of a sector may cause a long period of heightened regulatory uncertainty. From the moment liberalization is proposed and during the following years until the new market structure has taken effect, there is significant uncertainty about the conditions under which new investments in generation capacity will operate. This uncertainty continues for some time after liberalization, until the new system is fine-tuned and the market parties learn the dynamics of the new system. As a result, a decade

may pass before the market has recovered some form of routine. During this period, higher uncertainty constitutes a significant additional investment risk and, therefore, a reason to invest less.

A second lesson is that in electricity markets with a low demand price-elasticity, even generating companies with a small market share have market power during periods of scarce supply. The reason is that in such a situation it is possible to increase prices substantially by withholding only a small amount of capacity, so even relatively small generators have an incentive to withhold capacity. This means that the abuse of market power that was observed in California was not just an product of the market rules in California but is a phenomenon that may occur in any electricity market with low demand price-elasticity during periods of scarce supply. Forward contracts reduce the incentive to generating companies to withhold power and provide a hedge against price risk to retail companies.

Finally, the crisis demonstrated the need for consumers to be involved in the market. The fact that consumers did not respond to high wholesale prices (because consumer prices were fixed) contributed substantially to the crisis, as demand was not reduced when supply was tight. The high response to the call for voluntary conservation and load-shifting to off-peak hours during the crisis demonstrated that demand price-elasticity can be significant. The next chapter will evaluate the meaning of these lessons for other electricity systems and investigate the question of generation adequacy in more general terms.

5 The question of generation adequacy

In theory, periodic price spikes should provide optimal investment incentives in an energy-only market. However, several factors may cause the investment equilibrium to deviate from the social optimum. Combined with the high volatility of electricity prices, this creates a risk of investment cycles. The analysis is complicated by the fact that the optimal volume of generation capacity cannot easily be determined. In the presence of uncertainty, generating companies have an interest in erring on the side of too little generation capacity whereas society would rather over-invest, if erring is inevitable because the social costs of shortages due to too little generation capacity increase much faster than the costs of excess capacity. This divergence of interests combined with the high social cost of a prolonged period of scarce electricity generation capacity, gives reason to change the market structure to ensure future generation adequacy. This analysis leads to a set of criteria for evaluating policy intervention options.

5.1 Introduction

5.1.1 The question

In this chapter the issue of generation adequacy in liberalized electricity markets is analyzed. The question at hand is whether liberalized electricity markets tend to invest sufficiently, and in time, in generation capacity, so the probability of electricity shortages and the resulting service interruptions remains near the social optimum. There is no consensus in the scientific literature whether liberalized electricity markets can be expected to produce adequate capacity levels continuously. While there are some cases (most notably California) in which a liberalized market appears to produce insufficient investment in generation capacity, practical experience is too limited, and the available cases are too much convoluted by factors such as flawed market design or regulatory restrictions, to provide convincing empirical evidence. The social cost of capacity shortages is so high, however, that a thorough analysis of the issue is called for. If it

appears at all possible that electricity markets may develop periods of insufficient generation capacity, it is only prudent to review policy options.

The lack of scientific agreement on the issue is reflected in the different designs of electricity markets. Spain and several South American systems try to stimulate investment in generation capacity by providing capacity payments to generation in addition to their revenues from the sale of electricity (Vázquez et al., 2002). Three systems on the East Coast of the USA (PJM, the New York Power Pool and the New England Power Pool) use a system of capacity requirements to ensure a certain reserve margin (PJM Interconnection, L.L.C., 2003; see for an introduction Hobbs et al., 2001b). In response to the tightening supply of electricity, since early 2001 the Norwegian system operator tenders for operating reserves in contracts with a duration up to a year (Nilssen and Walther, 2001). Sweden created a mothball reserve around the same time (Lindqvist, 2001). In response to the shortages in June of 2003, Italy is considering implementation of a capacity mechanism as well (Fraser and Lo Passo, 2003).

Most other European systems, and California before the crisis, on the other hand, have no specific provisions to ensure capacity adequacy. Instead, they rely on the electricity market to provide investment incentives. They can be characterized as *energy-only markets*, as the (expected) price of electric energy is the only driver of capacity investment. This type of electricity market is the subject of this chapter. Experience with existing capacity mechanisms, as well as proposals for other capacity mechanisms, will be discussed in Chapter 6. Table 5-1 lists some electricity systems that have taken measures to stabilize generation investment.

Table 5-1: Experience with capacity mechanisms

Type of capacity mechanism:	**Implemented by:**
Fixed capacity payments	Spain, Argentina, Colombia
Dynamic capacity payments	England and Wales Pool
Long-term contracts for operating reserves	Norway
Strategic ('mothball') reserve	Sweden
Capacity requirement	PJM (USA, East Coast), New York Power Pool, New England Power Pool

5.1.2 Approach

A largely qualitative analysis of generation investment is presented in this chapter for the purpose of determining whether there is a need for policy intervention and, if so, what the specific goals should be. The leading question is whether energy-only markets can be expected to produce a socially acceptable outcome. Unless stated otherwise, the analysis in this chapter will therefore pertain to energy-only markets. Other types of market structures will be discussed in Chapter 6.

Starting point for this chapter is a literature review (Section 5.1.4). The theory of spot

pricing holds that competitive electricity markets provide a socially optimal outcome in both the short and the long term. From the fact that a number of mechanisms have been devised in order to stimulate investment in generation capacity in competitive markets, it may be concluded that this theory is not generally accepted to be fully applicable in practice. However, a comprehensive analysis of the reasons as to why a competitive electricity market may not provide a socially optimal outcome in the long run is lacking. This chapter intends to fill this void.

The lack of empirical data forces the analysis to focus upon the incentives which existing generating companies and newcomers to the market have to invest in generation capacity. The goal is to ascertain whether there is reason to believe that electricity markets may not produce an optimal result in the future. It is argued that the precautionary principle applies: if there is reason to believe that investment in generation capacity may not be sufficient, the high social cost of shortages is reason to consider alternative policy options.

Four aspects of the question of investment in generation capacity are analyzed. The analysis starts in Section 5.2 and 5.3 with a static analysis of the optimal investment equilibrium and factors that may disturb it. Section 5.4 is dedicated to the role of uncertainty regarding the optimal volume of generation capacity. Section 5.5 provides a dynamic analysis, that focuses upon the risk of the development of a business cycle. Section 5.6 discusses the role of market power, for instance the possibility that price spikes are manipulated like they were in California. Sections 5.7 and 5.8 discuss two additional issues: the first reviews the possible impact of technological changes upon the issue of generation adequacy, while the second analyzes the role of trade with neighboring systems. Section 5.9 discusses the public policy choices that are to be made. Section 5.10 presents the conclusions, including criteria for adjustments to the market design.

5.1.3 Some technical aspects of generation adequacy

The probability of shortages

All generating units have a probability greater than zero that they are not available at a certain time. Peak demand also develops in an unpredictable manner. Therefore, the probability that the available generation capacity temporarily is insufficient to meet demand can never be ruled out altogether. More generators reduce the probability of outages, *ceteris paribus*, but also cost more. Consequently, the socially optimal volume of generation capacity is not reached by minimizing the probability of outages but by keeping the total social costs (including the costs of the electricity supply industry) to a minimum. Section 5.4 discusses the role of uncertainty in the determination of the optimal volume of generation capacity and in investment decisions. A substantial literature exists with respect to reliability analysis of power systems (cf. Billinton and Allan, 1984; Billinton and Allan, 1992; Billinton et al., 1991.)

The load-duration curve

This chapter focuses on generation adequacy as defined in Section 2.5, meaning the

volume of available generation capacity. Load is commonly divided into base, medium and peak load, as illustrated in the sample load-duration curve in Figure 5.1. Generation capacity can be divided in a similar manner, as different types of plants are designed to serve the different load segments.

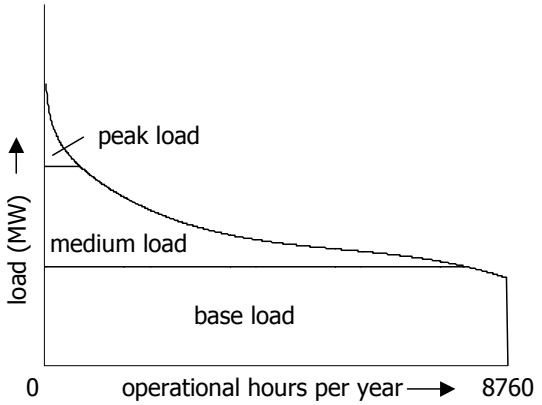


Figure 5.1: Sample load-duration curve

Along the X-axis, Figure 5.1 shows the hours in a given year that load reached the value indicated along the Y-axis. Rather than presenting these values chronologically, the load-duration curve ranks them by decreasing load. A load-duration curve therefore shows the number of hours that load was larger or smaller than a certain value. The load-duration curve changes from year to year, as demand grows and consumption patterns change. Consequently, exact load-duration curves for future years can only be estimated from historic load-duration data and load growth estimates. Unfortunately, the variability between annual load-duration curves is greatest for peak load, so this is the most difficult to forecast.

Load-duration curves can be used to estimate the operational hours of generators by ranking them in merit order along the Y-axis of the load-duration curve.[7] In addition to the error in the forecast of the load-duration curve, the forecast of the operational hours of a specific generator is affected by uncertainty about the availability of the generators that are higher in merit order. The lower their availability, the more hours the generator in question will be called upon (the lower it will be ranked along the Y-axis in Figure 5.1).

The power plants that serve base, medium and peak load have different technical and economic characteristics. The exact boundary between the categories is arbitrary, however. For the analysis of generation adequacy this distinction will be disregarded; consider only the total volume of generation capacity will be considered. Equally arbitrary is the distinction of a generation 'reserve' margin. A reserve margin can be

[7] This is true only for capacity-constrained systems, which are the focus of our analysis. Energy-constrained systems need to consider the energy content of their reservoirs and the expected inflow into the reservoirs.

defined as the difference between total available capacity and the highest system peak. However, as both functions are stochastic, predictions of a reserve margin have a stochastic nature as well. Marginal generators may move in and out of merit, and therefore out of and into the reserve, depending on demand and on the availability of generators higher in merit order. Therefore the perception of a fixed reserve margin, consisting of specific generating units, is misleading.

A more useful perspective is to consider all generating units as part of the same group, with their differences in marginal operating cost and ramp speed as the main determinants of their dispatch. If all generators are ranked by increasing marginal cost, their expected operating time decreases continuously but never reaches zero.[8] In this continuum, reserve units are simply those units with an expected operating time below a certain arbitrary value. While the expected operating time per year is greater than zero for all generating units, some of these units may not operate at all during some years.

The implication for the analysis is that one should not think of generating units as belonging to one or another category but rather as a continuous range of units which are dispatched by their owners based upon their economic merits, technical characteristics and availability. What used to be called reserve units are in fact nothing other than regular peaking units with a low number of expected operating hours. Generation adequacy is determined by the total available generation capacity in relation to peak demand.

Operating reserves versus capacity margin

A distinction should be drawn between operating reserves, which are necessary to maintain the physical, operational stability of electricity systems, and long-term reserves for the purpose of generation adequacy. Operating reserves typically are tiered, consisting of spinning reserves and reserves that are available within a certain time period of time (Billinton and Allan, 1984). They are used by the system operator to correct, from moment to moment, for deviations from the scheduled output of generators and from projected consumption.

The subject of this chapter is not these operating reserves – the use of which is not questioned – but the question of whether a capacity margin is required in addition to the volume of generation capacity that the market provides for the sake of securing electricity supply in the long term. The function of this capacity margin would be to provide a buffer in case peak demand is higher than projected or generator availability is lower than expected, or both. Due to the long lead time for new generation capacity, a capacity margin would not only improve system security in the face of operational contingencies but also make the system more robust with respect to unexpectedly high growth in demand. The capacity margin may be defined as the difference between total available generation capacity and peak load. The terms ‘reserves’ and ‘reserve capacity’, which often is used to describe the capacity margin, is somewhat confusing, as there is no

[8] Because there is a probability greater than zero for each generating unit that it is not available, there also is a probability greater than zero that even the last unit in the merit order will be called upon. This means that the expected operating time of this unit is greater than zero.

specific set of generating units that belong to a reserve as opposed to 'normal' generators. Therefore the term will be avoided here.

Data

Analyzing the issue of generation adequacy is complicated by a lack of relevant data. As mentioned in the introduction, there is currently a lack of empirical evidence to indicate whether competitive electricity markets provide a sufficient volume of generation capacity. Forecasting generation adequacy far enough in advance to be able to intervene before a shortage develops is nearly impossible as long as the lead time for new generation capacity remains several years. Therefore, it is difficult to determine whether the current investment pattern in competitive electricity markets is adequate. Even assessing whether the current volume of generation capacity is adequate may be difficult. Data regarding the availability of installed generation capacity is not necessarily available, and even an accurate tally of all installed capacity may be difficult to obtain. In the Netherlands, for example, 20 – 25% of generation capacity is in small, decentralized units, which are connected to distribution networks rather than to the transmission network. Data regarding installed capacity (for units larger than 10 MW) are being published since recently (TenneT, 2004). However, data on availability are not available for the Netherlands (Van Eck et al., 2002; EnergieNed, 2003).

The availability of generating units is determined by their technical reliability and the availability of primary energy but also by factors such as restrictions on waste heat, the availability of emissions credits and the linking of combined heat and power to other (industrial) processes. The Netherlands may be a difficult case; installed generation capacity may be easier to estimate in countries with a smaller share of decentralized units. A license requirement for supplying electricity to the grid may improve the availability of this information but creating a reliable data base for all the members of an interconnected system is a difficult task. In Europe, UCTE has taken this effort upon itself but appears to have overlooked the load served by decentralized generating units, which is not registered by the TSO (UCTE, 2002a).[9] This demonstrates the difficulty of obtaining accurate numbers in a field in which a few percentage points can be crucial. In North America, the NERC compiles this data and makes it available on its web site (North American Electric Reliability Council, 2004).

Regarding load, there also is a paucity of information regarding the development of load and the social cost of electricity service interruptions. The former can easily be solved by monitoring markets better. The social cost service interruptions – also known as the average value of lost load – is difficult to measure. (See the end of Section 5.2.4.) As this is an essential input in the calculation of the optimal volume of generation capacity, it makes this calculation itself uncertain, as will be seen in Section 5.4.2.

[9] The Union for the Co-ordination of Transmission of Electricity (UCTE) is an organization of transmission system operators in western continental, central and southern Europe. Comparison of their data with Van Eck et al. (2002) shows that the UCTE data likely only reflect load served through the transmission network and therefore disregards the 20-25% of decentralized power generation.

These difficulties in obtaining such basic data with any accuracy, in turn, underscore how much more difficult it is to obtain the necessary data for a quantitative analysis of the issue. The lack of data affects policy makers as well as market parties. Policy makers would like better data to monitor the development of the market (cf. Directive 2003/54/EC). A more transparent market would help also generating companies with their investment decisions, as will be argued later in this chapter.

5.1.4 Literature

A good introduction to the subject of security of supply, including transmission and fuel issues, is provided by Ocaña and Hariton (2002). They list a number of possible threats to the adequacy of electricity generation and briefly touch upon some solutions. They describe a number of case studies and note that generally, reserve margins are still high in Europe. However, they also show ample reserve margins in Norway and Sweden, where there currently are concerns about the ability to meet demand if rainfall is not above average (Nilssen and Walther, 2001). The UCTE (2003) is less optimistic about he development of reserve margins in Europe and warns for "a potential deficit in generation unless additional firm investment decisions are taken soon". In any case, the short history of liberalization provides insufficient experience to draw firm conclusions, given the fact that most systems started with excess capacity, which was reduced in the course of the first number of years. The fact that in nearly all cases the starting point for liberalization was a situation of excess capacity is probably the reason that the issue has not received much attention until now, as it was not immediately relevant.

Theory

The theory of spot pricing was introduced by Schweppe (1978). Caramanis, among others, further developed the theory and applied it to investment in generation capacity (Caramanis, 1982; Caramanis et al., 1982). They showed that under ideal conditions, electricity spot markets provide efficient outcomes in both the short and the long term. This theory stands; the question is whether it applies in practice, or whether real market conditions deviate too much from the ideal situation. The belief that unregulated markets in electricity generation can produce an optimal outcome in the long term is widely shared (cf. Shuttleworth, 1997; Hirst and Hadley, 1999; EnergieNed, 2002). Generally, this school of thought asserts that underinvestment would be caused by obstacles to the proper functioning of the market mechanism, such as price restrictions or construction permits. The correct course of action, in this view, is to improve the investment climate by eliminating all extraneous sources of risk, such as regulatory risk, and other obstacles to investment.

Stoft (2002) provides a thorough analysis of the relationship between peak load pricing and investment. He argues that occasional failure of electricity markets – defined as a situation in which the supply and demand functions do not intersect, so the market does not establish a price and demand needs to be curtailed – is inevitable if both supply and demand are inelastic and demand is volatile, conditions which are present in most electricity markets. This is a significant deviation from Schweppe and Caramanis, who assumed the presence of sufficient demand price-elasticity so service interruptions due to

supply shortages do not occur. Stoft shows that the social cost of market failure can be minimized by capping the electricity price at the average value of lost load. This will not prevent the occurrence of shortages with service interruptions but limits them to an economically efficient duration.[10] This analysis is briefly summarized in Section 5.2. Stoft (2002) presents two significant practical drawbacks to this market design. One is that the investment risk is high, which may easily lead to underinvestment if a market is not perfectly competitive. This point will be discussed in Section 5.3. The second issue is that the potential for high prices provides a strong incentive to withhold capacity. Section 5.6 discusses this point. Stoft therefore argues for a form of regulatory intervention to stabilize the generation market. (This option, called operating reserves pricing, will be discussed in Chapter 6.)

Borenstein and Holland (2002) show that even perfectly competitive markets will not produce an optimal volume of generation capacity if some consumers pay a flat rate for their electricity (so their demand is inelastic with respect to changes in wholesale prices) while others are exposed to real-time prices. The difference in prices paid by the different groups of consumers results in misallocation, the magnitude of which depends upon the differences in demand price-elasticity between the fixed-rate consumers and those on real-time pricing. This, in turn, leads to a sub-optimal volume of generation capacity, compared to what would have been optimal given the presence of a group of fixed-rate consumers. (The latter already is a second-best optimum, compared to the optimum in a market with full demand participation.) This finding is relevant because most markets have a mix of real-time prices and fixed rates. Borenstein and Holland show that in theory this imperfection can be corrected by adjusting the flat rate through a tax or subsidy (depending upon the situation).

Market failure

In a report to the State of Maryland, Hobbs et al. (2001c) argue why in practice energy-only markets can be expected to fail to maintain generation adequacy. Whereas Stoft focuses on the causes of the high investment risk in generation capacity and the resulting vulnerable investment equilibrium, Hobbs et al. focus on factors that may structurally impact the equilibrium. Hobbs et al. (2001c) note that the theory of spot pricing is based upon the following assumptions:

- Existence of 'real-time prices', meaning that both consumers and producers know the actual, momentary price. Absence of price distortions in the form of restrictions to prices, taxes or externalities.
- There is no market power.
- Generating companies have perfect knowledge of future prices and their stochastic distribution.

Violation of any of these assumptions, Hobbs et al. (2001c) argue, will lead to market

[10] This is a second-best optimum, as Borenstein and Holland (2002) also describe: the economic optimum given incomplete consumer participation in the market. The optimum is difficult to determine in practice because the social cost of unserved energy is difficult to determine. It depends, upon others, upon the time of day, the duration, the type of consumer, the frequency, whether the consumers were warned in advance, et cetera.

failure. They proceed to describe the different categories of market failure that may arise, an analysis that will be used in Section 5.3.

A number of other scholars agree that regulatory intervention is called for, despite the theoretical adequacy of the incentives provided by spot markets (Oren, 2000; Newbery, 2001; Vázquez et al., 2002). Oren and Newbery both consider the main reason to be the uncertainty that energy markets fully reflect scarcity rents, e.g. due to the last-minute imposition of price caps during price peaks. Vázquez et al. (2002) also cite risk aversion and market power as significant obstacles. Besser et al. (2002) go a step further, and argue that the price volatility in an energy-only market itself is unacceptable to consumers and regulators. These authors agree that regulatory intervention, the subject of Chapter 6, could provide a mechanism to ensure that prices remain within a socially acceptable range without deterring investment.

Doorman (2000) has written a dissertation on the question how to secure peak-load investment in restructured electricity markets. He considers risk aversion an important factor and notes that independent system operators generally lack the authority to take measures to secure generation adequacy. He cites the decreasing reserve margins in England and Wales, in Norway, and especially in California and Sweden as an indication that in restructured markets the likelihood of deficiencies is unacceptably high. The majority of his thesis is dedicated to analyzing solutions to this problem and, in particular, to presenting a new solution. This will be discussed in detail in Chapter 6.

Castro-Rodriguez et al. (2001) also arrive at the conclusion that energy-only markets invest too little but for different reasons. They assume that generators are an oligopoly with a strategy of under-investing because this is the only way to keep prices high enough to make a profit. Pérez-Arriaga (2001) concurs that in an energy-only market, an oligopoly will lead to a degree of underinvestment if there are entry barriers, which there are.

An entirely different line of argument is provided by Jaffe and Felder (1996), who reason that generation adequacy is a public good and will therefore be under-provided by a competitive market. In brief, they argue that a generator that does not operate does not earn any revenues, whereas it still provides a social benefit in the form of improved reliability. Jaffe and Felder argue that generation therefore is under-valued, so underinvestment is a likely result. Abbott (2001) and Besser et al. (2002) concur. Section 5.2.3 evaluates this argument.

Investment cycles

A final issue was first brought forward by Ford (1999), who forecasted the possible development of a generation construction cycle in California about a year before the crisis developed. Ford uses a computer simulation in which the cause of the cycle is the time lag between the occurrence of high prices and the availability of new generation capacity. In times of excess capacity, the electricity price is too low to attract new investment. Due to the low elasticity of supply and demand and the volatility of demand (factors which Stoft (2002) lists), prices rise quickly when demand has exhausted the available generation resources. During these price spikes, generators make large profits

which attract new investment but this only becomes available after a delay. As a result, the electricity price tends to oscillate, in some cases with increasingly high peaks, depending upon, among others, the rate of demand growth, the cost of new capacity and the time it takes to realize new generation capacity. Borenstein (2001) agrees that the extremely low elasticity of both supply and demand cause price volatility, which is exacerbated by the capital-intensiveness of generation, with boom-and-bust cycles as a likely outcome.

The tendency towards investment cycles is corroborated by Skantze and Ilic (2001), who also base their conclusions upon a computer simulation. They contend, however, that if mature forward markets exist that contain sufficient information, the cycles could be dampened and investment would be optimal (see also Visudhiphan et al., 2001). Stoft (2002) also holds the opinion that the existence of a feedback loop (in the form of higher prices when there is a need for more generation capacity) is not a sufficient condition to reach a market equilibrium, so investment cycles are a plausible scenario. He concurs with Skantze and Ilic that the delay in the response time of investors would not need to be a problem if the investors look far enough ahead into the future. However, Stoft argues that the long-term contracts will not cover enough time and will not contain enough information to engender optimal investment behavior. He points out that, due to the stochastic nature of the supply and demand functions, there always will be some periods (lasting a number of years) with lower prices than other periods. It requires a long time horizon (more than a decade) to be able to estimate correctly the average revenues from generation capacity. In practice, investors have a much shorter time horizon and they base their revenue expectations upon recent experience. As a result, they will tend to invest too little or too much, depending on recent market history, and contribute to the development of a business cycle. Henney (2004) recently provided an overview of the arguments why an investment cycle is to be expected in the England and Wales market. Most of his arguments apply to other energy-only markets as well.

5.2 Investment in a perfectly competitive market

5.2.1 Investment incentives in theory

The concept of a competitive electricity market is founded upon the theory of spot pricing of electricity. Caramanis et al. (1982) first described how spot pricing of electricity could be feasible and concluded that it would improve the economic efficiency of the electricity system. In a follow-up article, Caramanis (1982) describes how an electricity spot market not only leads to an efficient dispatch of generation in the short term but also leads to a socially optimal level of investment. The essence of the argument that Caramanis and his colleagues make is that conventional, neo-classical economic theory also applies to electricity generation. Thus, they were among the first to deny the conventional wisdom that electricity generation is part of a larger natural monopoly that encompasses the entire electricity production chain. Stoft (2002) provides a more detailed overview of the theory.

Even in an idealized, theoretical case, investment in electricity generation capacity is

different from investment in other sectors. Until now, no commercially viable way has been developed to store electricity, other than in hydro facilities. Because the network also does not store electricity, supply and demand must be in balance continuously. In a competitive market, this raises the question who will pay for the marginal generation unit, which operates only a small number of hours per year. Caramanis contends that in a perfect market, an investment equilibrium develops that is socially optimal. This can be seen as follows.

A generator's expected revenues in a certain year are determined by the number of hours that the generator operates in that year and the average electricity price during its operational hours. The generator will recover its costs when its revenues exceed the sum of its fixed and variable costs. For a peaking unit, the number of operational hours is small. Therefore, it will need high prices during its operational hours to recover its costs.

The theory that a spot market allocates available electricity resources efficiently is based, among others, upon the assumption that the demand for electricity exhibits sufficient short-term price-elasticity. If this is the case, supply scarcity, through high prices, leads to a reduction in demand, so outages do not occur. In the presence of sufficient demand price-elasticity, the market will therefore always clear, which means that scarce resources will be allocated efficiently (Schweppe 1978; Caramanis et al. 1982). Prices may rise to high peaks, however, in times of high demand: a portion of demand is characterized by a high value of lost load, while the supply function ends in an inelastic section when all available generation capacity is operating. These high prices allow peaking units to recover their cost.

If prices are high, generating companies invest in peaking units until the long-run average electricity price has decreased to the long-run average cost of a new generation unit. A balance develops between the average value of lost load, which is high for many consumers, and the high cost of maintaining peaking units to operate only a limited amount of time. So consumers influence the volume of generation capacity through their willingness to pay. Theoretically, the ensuing equilibrium is therefore socially optimal, as the cost of satisfying more demand would exceed its social value, while investing less would leave some customers unserved who would be willing to pay the price. This reasoning is similar to the neo-classical analysis of the long-run equilibrium between supply and demand. The only difference is that in electricity markets the demand curve shifts continuously, so generating companies are faced with daily variations in demand.

5.2.2 Low demand price-elasticity

In practice, the observed price-elasticity of demand is extremely low. In the remainder of the analysis, this therefore will be assumed to be the case. This is a crucial assumption: were demand price-elasticity significantly higher, electricity prices would be more stable and the need for random service interruptions would disappear, as will be seen below. A necessary condition for demand to be price-elastic is that consumers have access to real-time price information and that their bills reflect the time of day at which they use electricity (Hobbs, 2001c). To this end, final consumers would need to have real-time meters. This is currently not the case with a large proportion of consumers, especially

smaller ones. As their consumption is measured over periods of weeks or months, their bills can only reflect the average wholesale price during the billing period. Consequently, individual consumers do not save by avoiding consumption during peak times, so whatever price-elasticity exists cannot manifest itself. As a result, the observed price-elasticity may be significantly lower than the real price-elasticity of demand. There are multiple experiments aimed at increasing consumer price-elasticity but in most electricity systems their impact still is small (Nilssen and Walther, 2001; Roberts and Formby, 2001; Sæle and Grønli, 2001). See also Section 5.7. The intrinsic price-elasticity of demand may be low, however, because there is no readily available alternative for most applications of electricity.

Thus, the absence of real-time pricing, in a situation where supply and demand need to be balanced on a continuous basis, disturbs the feed-back loop between supply and demand. High spot prices do not lead to a reduction in demand according to consumers' willingness to pay. Most other mechanisms that aid the clearing of other markets, such as a delay in the delivery of the good, consumers switching to other goods, or higher prices leading to a reduction in demand, are not available in current electricity markets. This has significant consequences: wholesale electricity prices are highly volatile (more than in the theoretical model) and there is a probability of service interruptions.

A consequence of the high volatility of electricity prices is that investment in marginal peaking units is risky. Peaking units need to recover their costs during short price spikes, the frequency, duration and height of which all are highly uncertain in real markets. The investment risk is increased by the length of the time between the decision to build new capacity and the moment it becomes available. This time lag is caused by the construction time of new generation facilities, including the time required to obtain the necessary permits. To make socially optimal investment decisions, investors in peaking capacity would need to know the likelihood of price spikes and their expected height and duration. To this extent, they would need to know the stochastic distribution and expected growth of both demand, including exports, and the supply of electricity by their competitors, including imports into the system.

5.2.3 Generation capacity as a public good

The second consequence of the weak price mechanism, caused by the combination of the low demand price-elasticity and the need to balance supply and demand continuously, is that there is a risk that the market does not clear. There are periods when physically there is not enough generation capacity available to meet demand. When the price mechanism fails to ration demand, a different rationing method needs to be applied to manage shortages: controlled service interruptions, also known as rationing or rolling black-outs. From a perspective of economic efficiency, it would be best if service interruptions would be applied first to consumers with the lowest willingness to pay. However, in most electricity systems it is not possible to interrupt the service of specific consumers on short notice. As a consequence, service interruptions typically are somewhat random. The absence of economic criteria in the determination of who is to be interrupted means that they constitute a loss of economic efficiency.

If it is possible to interrupt or limit consumption easily on an individual basis, as would be the case for instance in a system with capacity subscriptions (Sections 6.7, 7.8 and 8.2.4), the issue is changed completely. Then it would be possible to use a price mechanism to ration demand. However, because the necessary infrastructure for this purpose is not available in most electricity systems, its will be presumed absent in the remainder of this chapter.

The chance of service interruptions is key to the existence of market failure in generation capacity, according to Jaffe and Felder (1996). The more generation capacity is available, the higher is the reliability of the supply of electricity. Therefore, they argue, the presence of generation capacity in excess of the capacity that is contracted by market parties ('reserve capacity') provides an additional benefit to all consumers of electricity in the form of higher reliability of service.[11] This benefit to all users of the system is a positive external effect of the provision of capacity, as the owner of the generation capacity cannot charge consumers for increasing the reliability of service. As the added reliability is non-excludable and non-rival (the reliability of service to all consumers increases), the generation capacity that is not contracted but is stand-by can be characterized as a public good.

The public good character of generation capacity is caused by the fact that all electricity is transported over a single network, as a result of which consumers cannot distinguish between the reliability of different generating companies, and the fact that in current systems electricity service cannot easily be interrupted on an individual basis. All generators connected to a network together contribute to 'the' reliability of that network; all consumers connected to a certain network experience the same level of reliability of service.[12] Because part of the socially optimal amount of generation capacity is a public good, liberalized electricity markets will tend towards an equilibrium volume of installed generation capacity that is lower than the social optimum. This analysis is corroborated by Pérez-Arriaga and Meseguer (1997), who consider generators to deliver three distinct products: energy, operating reserves and capacity reserves. When generators are not paid for their capacity reserves, they provide an external benefit.[13] A similar argument applies to the withdrawal of load: when a consumer reduces his load, system demand goes down and the probability of a shortage decreases. Withdrawal of load and the provision of additional capacity have the same positive external effect: they both increase system reliability.

It should be noted that this public good character of generation adequacy is a product of the current arrangements in most systems, which limit the possibilities to allocate scarce generation capacity. If these were improved, the public good character could be eliminated. One option is to limit individual consumers' electricity consumption during shortages to a level that was previously agreed upon ('capacity subscriptions', see

[11] For the sake of the discussion in this section, we will set aside the objections against the use of the term 'reserve' capacity which was mentioned in Section 5.1.3.
[12] This is true only to the extent that the reliability of electricity service is determined by generation; in many systems, network failures cause the majority of service interruptions.
[13] In some markets, the system operator pays for capacity reserves, as we will see in the next chapter.

Chapter 6). As this option requires technical changes to the system at the level of the individual customer, it may not be implemented easily, but it does mean that the public good character of generation adequacy is not an intrinsic characteristic of the system. Another, much favored option (cf. Hunt, 2002) is to install real-time meters with each consumer so they can react to current prices. This would be expected to significantly decrease price volatility and reduce the highest peaks in electricity consumption.

5.2.4 Value of lost load pricing: a second-best optimum

There is a counter-argument to this argument, however. While the existence of the externality is not denied, it can be shown that with some modifications the theory of spot pricing still holds, even if demand is assumed to be perfectly inelastic. Key is how the electricity price is set during a supply shortage (when there is no market equilibrium).

Figure 5.2 schematically shows the supply and demand curves in a market in which the price-elasticity of demand is insufficient to guarantee that supply and demand always match. In the short term, the supply function of electricity is fairly static; it is only influenced by the availability of generating units and it has a firm (perfectly price-inelastic) end. In the long term, the vertical part of the supply curve can be moved to the right by installing more generation capacity. The demand curve fluctuates continuously. Therefore Figure 5.2 shows a range of possible demand curves: the shaded area in the figure can be interpreted as an interval within which the demand curve can be expected to be with a certain likelihood, for instance a 95% confidence interval. As the figure indicates, there is a possibility that the supply and the demand functions do not intersect: then there is a supply shortage and demand needs to be rationed to maintain system stability.

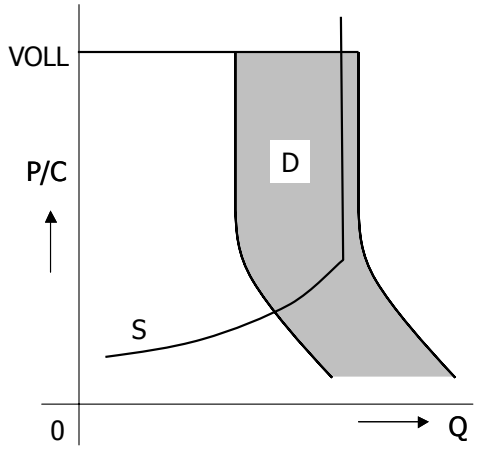


Figure 5.2: Insufficient demand price-elasticity means there is a possibility that supply and demand do not match

When the supply and demand functions do not intersect, the market does not clear and there is no equilibrium price. The consequences depend upon whether consumers are free to select their supply (retail) company or not. When there is a shortage, this means that the supply companies have insufficient contracts with generators to cover the consumption of their customers. In decentralized systems with full retail competition, the system operator requires supply companies to make up the difference between the electricity that they have sold and the electricity they have purchased in a balancing market. The requirement to purchase means that the supply companies have an inelastic demand function in the balancing market. Due to the shortage of electricity, there is no theoretical limit to the electricity prices. Therefore, it is necessary to implement a price cap to protect consumers against overcharging, (e.g. Ford, 1999; Hobbs et al., 2001b; Stoft, 2002). When consumers are not involved in real-time price setting, they otherwise may find themselves paying more for electricity than their value of lost load. This is indicated in Figure 5.2 by the horizontal section of the demand curve at a price equal to the average value of lost load (VOLL).

In markets without retail competition, there still is a need for a maximum price. It is true that when supply companies have regional monopolies, they may be able to interrupt service to certain groups of customers and thus adjust their demand to the market price. However, it is questionable whether the supply companies properly reflect their customers' willingness to pay for electricity. On the one hand, supply companies could reduce their purchasing costs by curtailing service to consumers sooner than would be economically efficient, as the companies do not bear the immediate costs of supply interruptions. On the other hand, the political and social repercussions of service interruptions may be so large that the supply companies may go far to avoid them, paying more than their consumers would. As Hobbs et al. (2001b) put it: "the height of price spikes today reflects the unwillingness of ISOs or load-serving entities (LSEs) to stomach the political fallout of curtailment, rather than the willingness to pay of the marginal power user". Therefore, consumers may end up paying more than their actual willingness to pay (but they will only see this in the form of a higher average price over the entire billing period). This is an argument for implementing a price cap equal to the average value of lost load in this situation as well.

A price cap provides an opportunity to influence generators' revenues. The higher the maximum price, the higher the expected revenues during periods of scarcity, and therefore the larger the incentive to invest in peaking units. Thus, it should be possible to compensate the under-incentive to invest due to the public good character of reserve capacity, which was mentioned in the previous section, with a sufficiently high maximum price. Systems without a maximum price should, in theory, even provide an incentive for over-investment.

The price cap needs to be set carefully, as it impacts the attractiveness of investment in generation capacity. In theory, the price cap needs to equal the average value of lost load (VOLL) because at this price consumers should, on average, be indifferent whether they receive electricity or not. Stoft (2002) shows that in a perfectly competitive market, a price cap equal to the average value of lost load results in an optimal level of investment in generation capacity. Investment in generation capacity takes place up to the point

where the marginal cost of generation capacity per unit of electricity produced is equal to the average value of lost load. In this equilibrium, social cost is minimized, as the cost of marginally more generation capacity would not be offset by the associated marginal benefit of fewer outages. Therefore, the theory of spot pricing still is valid, even if demand is fully inelastic. This version of spot pricing sometimes is called value-of-lost-load-pricing, as setting the price cap equal to the average value of lost load is a central element in the design of the market (Stoft, 2002).

The equilibrium achieved by value-of-lost-load-pricing is a second-best optimum, given the absence of sufficient involvement of the demand side. Capping the electricity price at the average value of lost load when the value of lost load varies significantly between consumers causes significant welfare losses. When load is shed, some consumers would have preferred to pay more to maintain service, whereas others, still being served, would rather have reduced consumption than paid a price equal to the average value of lost load, had they had the option. Therefore service interruptions cause a loss of economic efficiency compared to a market in which demand price-elasticity is sufficient to avoid the need for load shedding. Further inefficiencies are caused by temporal variations in the value of lost load of many consumers, and the fact that the cost of outages may not be constant with increasing duration of the outages and may depend upon the warning time which is given to consumers. These factors also make it notoriously difficult to estimate the average value of lost load (cf. Kariuki and Allen, 1996a, Kariuki and Allen, 1996b, Goel and Billinton, 1997, Ajodhia et al., 2002). This is a fundamental weakness of VOLL-pricing, as a wrong price cap would cause the market equilibrium to deviate from the optimum.

5.2.5 Summary

The argument made in this section is summarized in Figure 5.3. There are three factors that contribute to price volatility and the possibility of supply interruptions due to a shortage of generation capacity in the current structure of the electricity system:

- the absence of real-time pricing possibilities,
- the absence of commercially viable storage facilities for electricity, and
- the fact that any physical shortage of electricity threatens system stability.

These three factors cause electricity prices to be highly volatile. This, in turn, makes investment in peaking units risky, as their revenues depend upon price spikes of which the frequency, height and duration all are uncertain.

A second effect of the low price-elasticity of demand and the fact that supply and demand always need to be in balance is that there is a possibility of supply interruptions due to a shortage of generation capacity. If consumers do not react to real-time prices, it may be possible that demand exceeds supply simply because consumers do not sufficiently reduce their demand in response to high prices. To protect consumers, a price cap equal to the average value of lost load is required.

The conclusion of the theoretical case presented in Section 5.2.1, that an unregulated market will tend to produce an optimal level of available capacity, still appears to hold. However, the investment equilibrium appears to be quite vulnerable due to the risk

associated with investment in peaking capacity and the difficulty in determining the average value of lost load. The economic viability of peaking plants is easily influenced by many factors, as will be seen in the next section. As a result, the investment optimum may easily be shifted away from the social optimum.

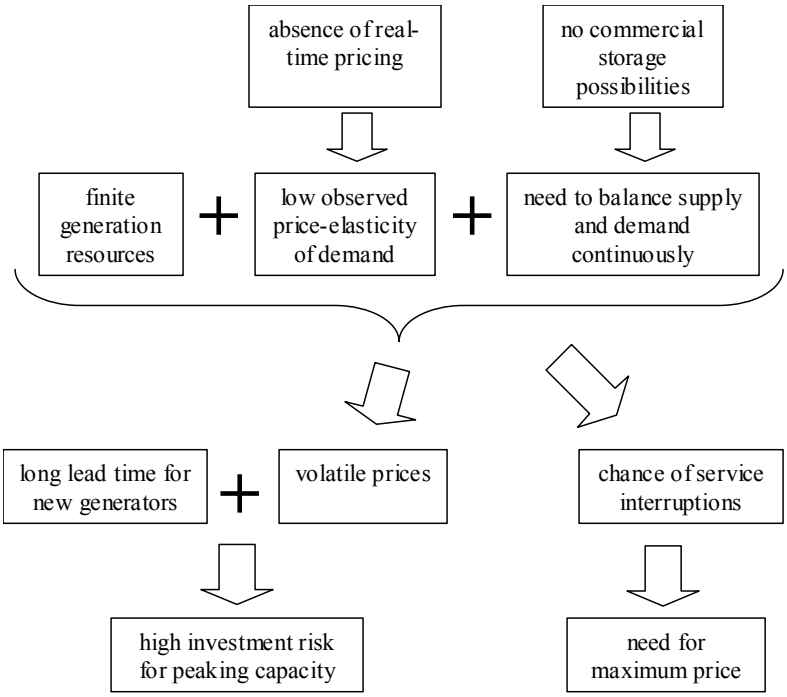


Figure 5.3: Overview of the argument

5.3 Factors influencing the investment equilibrium

Now a number of factors that may disturb the investment optimum will be discussed. The following factors may reduce the equilibrium volume of generation capacity (based, in part, upon Hobbs et al., 2001c):

- price restrictions,
- imperfect information,
- regulatory uncertainty,
- regulatory restrictions on investment,
- risk-averse behavior by investors,
- uncertainty regarding input markets, and
- externalities in the generating market.

5.3.1 Price restrictions

The fact that a price cap may be required to protect consumers against overcharging in times of scarcity represents a risk because the optimal level of the price cap is difficult to determine. While the theory is clear that the maximum price needs to equal the average value of lost load, there are many methods, with widely varying outcomes, to measure the average value of lost load (cf. Kariuki and Allen, 1996a, Kariuki and Allen, 1996b, Goel and Billinton, 1997, Ajodhia et al., 2002). Estimates range as high as € 15/kWh (€ 15,000/MWh) (Willis and Garrod, 1997) but other estimates are an order of magnitude lower (cf. Australian Competition and Consumer Commission, 2000). The cost of erring is high. Absent compensating measures, a price cap that is below the average value of lost load should be expected to result in a sub-optimal level of investment in generation capacity. Too high a price cap would lead to excessive wealth transfers from consumers to generating companies.

5.3.2 Imperfect information

Producers lack the information needed for socially optimal investment decisions (Hobbs et al., 2001c). On the supply side, information deficiencies increase the investment risk and thus lead to lower equilibrium volumes of installed capacity. In order to calculate the probability that their peak units will operate and to calculate the expected return on investment, generating companies need to know both the stochastic distribution of the demand function (so they know the distribution of the frequency, duration and height of price spikes) and the expected development of total available capacity (Hobbs et al., 2001a). The exact characteristics of the demand function are difficult to estimate, especially in newly liberalized markets for which no long time sequences of empirical data are available. Moreover, the basic characteristics of demand change over time (for instance due to the introduction of new technologies) which reduces the validity of demand functions built upon historical data.

5.3.3 Regulatory uncertainty

The willingness to invest is adversely impacted by regulatory uncertainty because it increases investment risk. Oren (2000) and Newbery (2001) among others, consider this one of the main factors leading to inadequate generation capacity. Regulatory uncertainty can be considered as a negative externality associated with changes in public policy. Especially in newly liberalized markets (which most electricity markets are), regulatory uncertainty can be a significant factor. Some examples of sources of regulatory uncertainty are the following:

- fine-tuning the market design,
- political intervention,
- changes in input markets,
- changes to the regulatory conditions for the market, and
- network expansion.

Fine-tuning the market design

When liberalizing a large and complex market like electricity, where many countries and interests are involved, it appears inevitable that the market design needs to be adjusted as our understanding of its dynamics evolves. However, this process of fine-tuning creates significant regulatory uncertainty for a long period of time. This runs counter to the goal of creating stable market conditions. The ensuing uncertainty undermines investment incentives, so one of the main goals for the system, reliability of service, is compromised. Thus we come upon a fundamental dilemma regarding the liberalization of a complex industry like the electricity sector: it is impossible to establish perfect market rules at the outset of liberalization but adjusting the rules along the way creates regulatory uncertainty, which undermines long-term system development.

For example, the first Electricity Directive of the EU (Directive 96/92/EC) was adopted in 1996, after at least five years of discussion. Seven years later, a new Electricity Directive was adopted (Directive 2003/54/EC). This, however, is not the end of the period of regulatory uncertainty, as the member states need time to implement the Directive and have considerable leeway in its interpretation. The deadline for liberalizing small consumers is 2007, which means that the transition to a liberalized market will have taken at least 15 years. This is close to half the technical life span of a generator, which means that a substantial part of the generating stock should have been renewed during this period. Fortunately, disincentives for investment during the transition phase are at least partly compensated by excess capacity that was present at the outset of liberalization.

Political intervention during a shortage

Section 5.2.4 described that prices occasionally need to rise to the average value of lost load in order to provide an efficient investment incentive in an energy-only market. However, most electricity systems started liberalization with ample capacity, so such high price spikes have not yet occurred in most systems. This transition effect raises the appearance that the political promise at the outset of liberalization, that the same or better service would be provided at lower prices, indeed is being met. If, after the initial excess capacity has disappeared, a period develops in which prices are many times higher than their historical levels, consumers may consider this a failure of liberalization and demand intervention, for instance by imposing a low price cap. This occurred in San Diego at the beginning of the California crisis, where even a brief period of high consumer prices proved politically unacceptable (Liedtke, 2000).

The political risk associated with extremely high electricity prices, whether these are economically efficient or not, means for investors that there is a risk that government intervenes and lowers the maximum price during a price spike. Hence price volatility itself brings about regulatory risk, at least until sufficient experience has been gained with liberalized markets so investors know whether they should expect political or regulatory intervention or not (Oren, 2000; Newbery, 2001).

The call for political intervention will be reinforced by suspicions of capacity withholding. The experience in California showed the public that generating companies

may have both a motive and the opportunity for price manipulation. The presence of high prices will therefore arouse suspicion, whether the generating companies really manipulate the market prices or not. This will increase the pressure upon politicians to intervene and impose a low maximum price. Therefore a system that relies upon price spikes to signal the need for investment may ultimately be politically unstable.

Changes in input markets

Natural gas is one of the main inputs in the production of electricity, which means that the current restructuring of the European natural gas market creates additional regulatory uncertainty for the electricity sector. The slower the restructuring process of the gas sector, the longer this uncertainty will last. Most notably the development of the gas transport tariff system, including charges for flexibility and imbalance penalties, is uncertain. This has a considerable impact on a business plan involving today's state-of-the-art gas-fueled generators.

Changes to the regulatory conditions for the market

There is uncertainty about future environmental rules, such as cooling water regulations or the effects of the recently adopted CO_2 emissions trading scheme (EC, 2002). A number of European countries have the intention to decommission their nuclear facilities. Given the long time schedules, the question is whether later governments may reverse this policy.

Network expansion

The European Union has as a goal to expand interconnector capacity, which would significantly alter market dynamics. Increasing transmission capacity may improve the competitiveness of the markets (Borenstein et al., 2000). However, in many cases environmental constraints or local opposition form an obstacle to construction of new power lines.

5.3.4 Regulatory restrictions on investment

Obtaining the necessary permits for the construction and operation of generating plants may present another obstacle to investment. While the social benefits of a proper licensing process are not disputed here, it should be taken into account that there also may be negative side-effects. First, the permitting process can be lengthy, which increases the response time of generating companies to an increase in demand. Especially in a situation of incomplete information about the future development of supply and demand, this may contribute to investment risk.

In an extreme case, high regulatory barriers may contribute directly to a shortfall in generation capacity. In Norway, the question of building a gas-fired power plant was so contentious that a government fell over the issue (Overbye, 2000). The new government decided to proceed with the construction of the plant. Had the political vote been otherwise, however, new construction in Norway would have been made quite difficult. Regulatory restrictions may not always be as explicit as in the case of Norway. For

instance the application of environmental standards that are reasonable for base-load plants may be too costly for peaking units. As the latter only operate a limited number of hours per year, the pressure to reduce capital cost is higher, while their environmental impact is smaller than in the case of base-load plants.

A final effect of permitting requirements may be to raise the barrier for new entrants to the market. Incumbents may be able to construct new plant at existing sites, for instance at the location of decommissioned old plant. This already has the advantage of having the infrastructures for electricity, fuel and cooling water present. Permitting requirements may have the effect of further discouraging greenfield development of new plant. While this may be desirable from the point of view of land use planning, the effect of stimulating oligopolistic behavior should not be disregarded.

5.3.5 Risk aversion

The theoretical approach of Section 5.2.1 assumes that generating companies behave in a risk-neutral manner with respect to investment. This is not necessarily the case, especially when many risks themselves are not well understood. Given the many non-quantifiable risks in a liberalized electricity market, it is not unlikely that investors in generation capacity choose a risk-averse strategy (Vázquez et al., 2002). If all investors do so, none of them lose market share so the penalty is limited to a loss of sales during periods of supply shortage. However, this loss of volume is small, compared to overall production of electricity, and is likely to be more than compensated by the high prices that develop during a period of supply shortage. Therefore a collective strategy of risk-averse investment behavior is beneficial to the generating companies, as long as this does not attract newcomers to the market. Such a risk-averse investment strategy would lead to less installed capacity than would be socially optimal. Section 5.4 will further develop the issues of uncertainty and risk.

5.3.6 Uncertainty regarding input markets

Disturbances in the markets for inputs for electricity generation may impact electricity markets. The obvious example is fuel markets. Especially electricity systems with a high dependence upon a single fuel are vulnerable to disturbances of the fuel supply. The risk consists not only of disruptions of the production of primary energy sources but also of disruptions of the supply infrastructure. In addition to physical disruptions, strategic manipulation of infrastructures, such as withholding of pipeline capacity, may threaten reliability. In California, this was one of the secondary causes of the crisis (FERC, 2002c).

Increasingly, emissions credits markets are used to reduce the environmental impact of electricity generation, such as the NO_x trading system in California and the planned CO_2 credit system in the European Union. These are artificial markets in which total supply is controlled by government. While they should provide a flexible and economically efficient system to reduce emissions, a poorly designed emissions market may interfere with the electricity market. A firm limit to the supply of emissions credits, for instance, may exacerbate an electricity shortage. Therefore emissions trading schemes should have

sufficient opportunities for banking and borrowing credits.[14] Moreover, there appears to be a risk that the emissions permits market is used to manipulate electricity prices (Kolstad and Wolak, 2003).

The risks from disruptions of input markets is two-fold. First, severe disruptions of input markets may jeopardize the reliability of the electricity system. More important for the issue of generation adequacy, however, is the fact that instability of input markets increases investment risk.

5.3.7 Externalities in the generating market

In all markets, the presence of externalities, positive or negative, causes the market equilibrium to deviate from the social optimum. Clearly it is important for the achievement of a socially optimal volume of generation capacity that price signals are not distorted by the presence of externalities. The main externalities associated with the production of electricity are negative environmental externalities, in the case of fossil fuels, and safety and waste issues in the case of nuclear power. Ignoring these external costs could lead to a higher consumption of electricity than would be socially optimal. The extent of these externalities is correlated to fuel consumption. As peaking units produce little electricity, the social cost of fuel-related externalities is small in proportion to the capital cost. Therefore the presence of negative environmental externalities will likely not shift the optimal volume of capacity by much. Moreover, the issue of externalities is not specific to the generation capacity market, and a wide body of literature exists about how to mitigate externalities. (Cf. Pearce and Turner (1990) for environmental externalities.) Therefore the issue will not be discussed here any further.

As mentioned in the introduction to this chapter, Jaffe and Felder (1996) raise a different issue, namely that generation adequacy, or reliability, as they call it, itself is a positive externality associated with the operation of generation facilities. This issue was discussed in Section 5.2.3.

5.3.8 Overview of the argument

Section 5.2 described how, in theory, an unregulated electricity market should provide an optimal volume of generation capacity but that the optimal investment equilibrium is vulnerable. The current section identified a number of factors that tend to discourage investment, often because they increase investment risk. A significant number of these factors appear to be unavoidable in real markets. For instance, it will probably never be possible to provide investors with all the information regarding future supply and demand that they desire. Similarly, the presence of regulatory uncertainty is inevitable during the long transition period to a liberalized market, not only for electricity, but also for related markets such as natural gas and CO_2 emissions permits. Therefore it appears likely that

[14] The recently adopted proposal for a tradable CO_2 emissions scheme in the European Union provides in banking, but not in borrowing. An objection against borrowing is that it shifts compliance to a later date. However, the California case showed that an acute shortage of emissions credits may aggravate a power crisis.

energy-only markets will tend to provide less generation capacity than is socially optimal. Figure 5.4 provides an overview of the argument made so far.

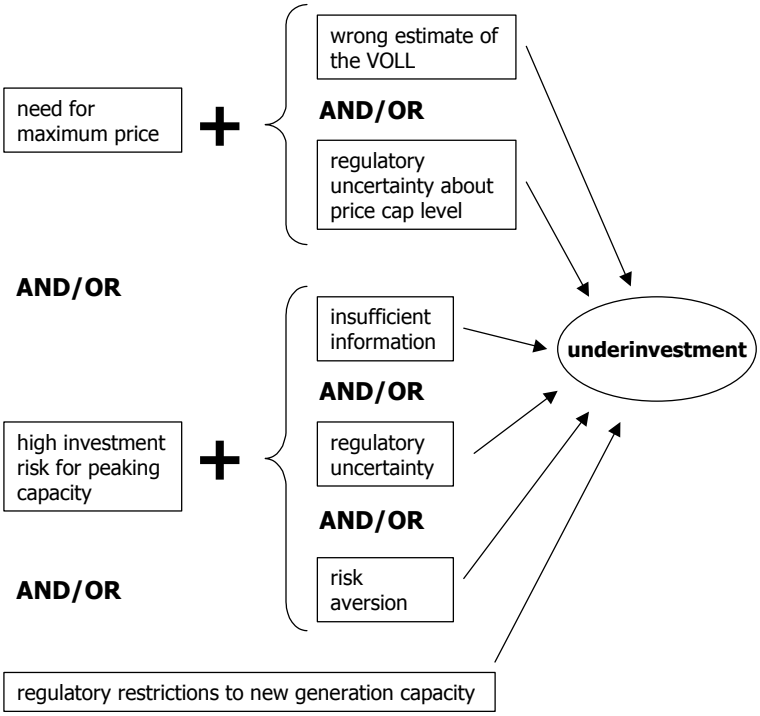


Figure 5.4: Factors in energy-only markets that may lead to insufficient investment in generation capacity

5.4 Investment and risk

5.4.1 Introduction

Section 5.2 argued that the investment equilibrium in an energy-only market, in which periodic price spikes signal the need for peaking capacity, is easily disturbed. Because the marginal peaking units need to recover their costs entirely during price spikes, operating these units involves a significant market risk. Errors in the forecasts of the height, duration and frequency of these spikes may have a significant impact upon investment behavior, and thus in the long run upon the reliability of electricity supply. Section 5.3 argued that these errors are easily made, as neither the long-term demand function nor the long-term supply function are well known, neither their average values nor their probabilistic distribution functions. Moreover, the height, duration and frequency of the price spikes themselves may readily be affected by a number of factors. In this section the impact of uncertainty and risk upon the decisions of generating companies and consumers

is investigated.

The analysis is limited to the question of the quantity of generation capacity. It is assumed that the generating companies maintain a diverse enough portfolio of generation capacity that they can follow the daily swings in demand adequate. It is further assumed that the system operator succeeds in ensuring a sufficient volume of operating reserves that are flexible enough to maintain operational stability of the system. Therefore only the total volume of generation capacity is of concern.

The life cycle of generation capacity is long, compared to the speed with which society and technology develop. As a result, it is unlikely that the generation market ever achieves a long-term equilibrium, in which the installed generation capacity is optimal for meeting society's needs. Rather, the inertia of the system will probably to cause it to always be moving towards, but never reaching, a continuously shifting equilibrium.

In this light, the question is not so much which volume of generation capacity is exactly socially optimal but rather how the expected social losses from erring can be minimized. This section analyzes the role of uncertainty in maintaining generation adequacy. Before starting, the concept of the optimal volume of generation capacity needs to be developed further. Section 5.4.2 describes how in theory (given sufficient information) the optimal volume of generation capacity can be determined. Section 5.4.3 discusses the implication of the stochastic nature of volume of available generation capacity. 5.4.4 argues that, in the presence of uncertainty, risk is distributed asymmetrically around the investment optimum. Consequently it is rational for society to create a larger volume of generation capacity than is estimated to be optimal. Section 5.4.5 discusses the perspective of the generating companies.

5.4.2 The optimal volume of available capacity[15]

Section 3.6.4 provided a brief introduction to the issue of system optimization. In this and the following sections the optimal volume of generation capacity will be analyzed. In theory (leaving aside demand growth and uncertainty), the optimal volume of generation capacity is found as a trade-off between installing more generation capacity, which reduces the cost of outages, and the cost of this capacity. The goal is to maximize the net social benefit of electricity service B_n (measured in monetary units, for instance €/y) as a function of available generation capacity q (measured in MW). B_n is equal to the sum of consumer surplus and producer surplus. B_n is equal to the total benefit of electricity B_s if all demand were served, minus the cost of outages C_o (€/y) minus the cost of generation C_g (€/y):

$$B_n = B_s - C_o - C_g \qquad (5.1)$$

[15] This section summarizes some pertinent parts of reliability theory. For a more extensive analysis, see for instance Billinton et al. (1991), Billinton and Allan (1992) and Kling (1998). Jonnavithula and Billinton (1998) describe a method for calculating the minimum of the sum of the cost of generation and the cost of outages.
The author would like to thank Hamilcar Knops for his support in developing the argument in this section.

The total benefit of electricity service B_s is difficult to measure. In theory, it is the integral of the demand curve but the precise demand function is not observable in current markets (Stoft, 2002).[16] This does not matter, however, because to know the optimal volume of generation capacity, only the derivative of (5.1) with respect the volume of available generation capacity q (MW) needs to be calculated. If demand is assumed to be fully price-inelastic, the total benefit of electricity B_s is not related to the volume of available capacity. The other two terms on the right-hand side of (5.1) are a function of q.

The cost of outages C_o is the product of the average value of lost load V_{ll} (€/MWh) and the volume of energy that is not served $E_{ns}(q)$ (MWh/y):

$$C_o = V_{ll} \cdot E_{ns}(q) \qquad (5.2)$$

The average value of lost load will be assumed to be independent from the average duration of outages and therefore to be a constant value. Substituting (5.2) in (5.1) and deriving with respect to q:

$$\frac{dB_n(q)}{dq} = -V_{ll}\frac{dE_{ns}(q)}{dq} - \frac{dC_g(q)}{dq} \qquad (5.3)$$

The units of $dB_n(q)/dq$ and dC_g/dq are €/MW per year and of $dE_{ns}(q)/dq$ are h/y.

Near the optimal volume of available capacity, that is, for values of q close to peak demand, the marginal cost of generation dC_g/dq is mainly determined by the fixed cost of peaking units C_f, measured in €/MW per year. (This assumption is commonly made, among others by Stoft (2002).) Then equation (5.3) becomes:

$$\frac{dB_n(q)}{dq} = -V_{ll}\frac{dE_{ns}(q)}{dq} - C_f \qquad (5.4)$$

To find the optimal volume of available generation capacity, (5.4) needs to be set equal to zero. Then:

$$\frac{dE_{ns}(q)}{dq} = -\frac{C_f}{V_{ll}} \qquad (5.5)$$

As the load-duration curve is assumed to be known, now the relationship can be established between the volume of unserved energy as a function of available generation capacity $E_{ns}(q)$ and the load-duration curve. The load-duration curve shows the volume of demand as a function of the number of hours per year. For our purposes, the inverse relationship is more useful. Figure 5.5 shows the inverse of the sample load-duration curve that was presented in Section 5.1.2. The function $f_d(q)$ is defined as the inverted load-duration function: it indicates the number of hours per year that load equals a certain

[16] Note that B_s is different from consumer surplus, which is the integral of the difference between the demand curve as a function of output and price.

volume q.[17] The units of $f_d(q)$ are hours per year. First the relation between $E_{ns}(q)$ and $f_d(q)$ is determined.

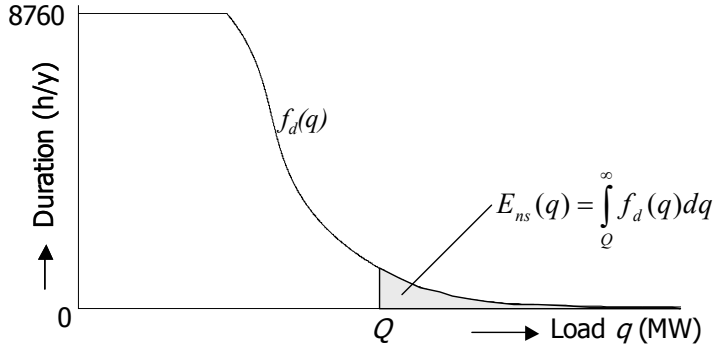


Figure 5.5: Inverted load-duration curve

If available generation capacity is equal to Q, then the expected volume of unserved energy is equal to:

$$E_{ns}(q) = \int_Q^\infty f_d(q)dq \qquad (5.6)$$

The load-duration curve provides the relationship between available capacity and duration, not energy. Therefore the optimal average amount of time per year that there is insufficient generation capacity available to meet demand should be calculated. This is also known as the loss of load expectation (LOLE). To find the optimal LOLE, (5.6) is derived with respect to Q:

$$\frac{dE_{ns}(Q)}{dQ} = \frac{d}{dQ}\left[\int_Q^\infty f_d(q)dq\right] \qquad (5.7)$$

If $F_d(q)$ is a primitive of $f_d(q)$, then

$$\frac{dE_{ns}(Q)}{dQ} = \frac{d}{dQ}\left[F_d(q)\right]_Q^\infty = \frac{d}{dQ}\left[F_d(\infty) - F_d(Q)\right]$$

[17] Note the load-duration curve actually only is a collection of data points. A mathematical function can only approximate this series of data points. For the present analysis, we will assume that we know a function $f_d(q)$ that is a close approximation of the data points.

Example 5.1: The optimal loss of load expectation

If the fixed costs of a peaking plant are 40,000 €/MWh per year (Newbery et al., 2003) and the average value of lost load is 8,600 €/MWh, as was recently established for the Netherlands (Bijvoet et al., 2003), the optimal loss of load expectation can be determined with equation (5.9). In this case, it would be 40,000/8,600=4.7 hours per year. This is the optimal average duration per year that not all demand in the system can be satisfied due to a scarcity of generation capacity. Outages due to network service interruptions are not included in this analysis. Clearly, this outcome is highly sensitive to the estimate of the average value of lost load and the estimated cost of peaking (reserve) capacity.

The same result also is obtained from the perspective of generating companies. The marginal generator (the generator with the highest variable costs in the system) would expect to earn a price equal to the average value of lost load during the few hours that it operates. At a price of 8,600 €/MWh, the last unit would need to operate 40,000/8,600=4.7 hours per year in order to recover its cost. Assuming that the size of the plant is small relative to the total volume of demand, the expected loss of load expectation also is 4.7 hours per year. This would be the economically efficient loss of load expectation because the cost of more generation capacity would be less than the benefit of reduced outages.

The amount of time that consumers can expect to be without service is much smaller than the loss of load expectation because only a small proportion of consumers are interrupted each time that load is shed. If it is assumed that during those 4.7 hours on average 2% of load must be interrupted (which was the maximum in the California crisis), the expected duration of service interruptions per consumers will be 2% of 4.7 hours, which is less than 6 minutes per year per customer. In comparison, the total average duration of power failures (mostly caused by distribution network problems) is about 25 minutes per customer per year in the Netherlands (EnergieNed, 2004).

To compare the revenues of the marginal plant to a peaking plant that is slightly higher in the merit order, consider a plant that has an expected operation time of 50 hours per year. This plant would only need to receive an average electricity price of 800 €/MWh during those hours to recover the same fixed costs. Clearly, estimates of the return on investment in this section of the market are highly sensitive to:

- the estimate of the average value of lost load
- the estimate of the load-duration curve
- the estimate of the generator's position in the merit order.

The latter two determine the expected number of operating hours of a generator.

$$= -\frac{d}{dQ} F_d(Q) = -f_d(Q) \qquad (5.8)$$

The optimal LOLE $f_d(q)^*$ can therefore be found by substituting (5.8) in (5.5):

$$f_d(q)^* = -\frac{dE_{ns}(q)}{dq} = \frac{C_f}{V_{ll}} \qquad (5.9)$$

Once the optimal loss of load expectation is known, the load-duration curve shows the associated optimal volume of available generation capacity q^*.[18]

Note that the optimal loss of load expectation that is determined with (5.9) is not equal to the expected amount of time that a consumer is without electricity. During the hours that supply is inadequate to meet all demand, the system operator needs to impose blackouts; however, each time only a small fraction of consumers is affected. Therefore the expected average duration of service interruptions for each consumer is only a small fraction of the loss of load expectation for the system, which was calculated with (5.9). The expected average duration of service interruptions per customer depends upon the load-duration curve. Example 5.1 illustrates the notion of the optimal loss of load expectation.

5.4.3 The optimal volume of installed capacity

Determining the optimal volume of available generation capacity with equation (5.9) and the load-duration curve appears simpler than it is. Example 5.1 already indicated some of the obstacles: the estimate of the optimal loss of load expectation is dependent upon a good estimate of the average value of lost load, which is difficult to obtain (cf. Kariuki and Allan, 1996a; Kariuki and Allan, 1996b; Willis and Garrod, 1997; Ajodhia et al., 2002). To find the optimal volume of available generation capacity, one also needs to know the load-duration curve. While good data keeping may reduce some of the uncertainties in estimating the load-duration curve, demand growth and possible shifts in consumption patterns will continue to create uncertainty. The future load-duration curve is influenced by all the factors that influence demand: weather, the general economy, the introduction of new consumer appliances, et cetera.

The most fundamental uncertainty, however, stems from the fact that the availability of generators is uncertain. As a result, the volume of available generation capacity is a probabilistic function of the total volume of installed generation capacity. To determine the distribution function of the availability of generation capacity in a system, one would need to know the stochastic distribution of the availability of each unit in the system (cf. Billinton and Allan, 1992). In the USA the North American Reliability Council (2004) collects this data in the General Availability Data System (GADS). In Europe, similar data is not available to the public, as far as the author knows. The availability of a generating unit may be correlated to the availability of other units, for instance because they are both affected by cooling water limitations. To complicate matters further, generating companies may be able to influence the availability of their generators to a

[18] This result is the same as the result that Stoft (2002) finds (Result 2-3.1). We have used slightly different units and notation than Stoft does: he uses $D(q)$ for the duration of load shedding, which is dimensionless, but can be interpreted as a fraction of time. The units of $f_d(q)$ are hours/year. Stoft disregards the stochastic nature of the availability of generation capacity, which is the subject of the next sections.

degree, for instance with their maintenance schemes.

Another route is to estimate available generation capacity empirically for the whole system. This is more feasible but the results will be less accurate because the estimate would be based upon a time series, while the generating stock and the dispatch pattern change from year to year. Generally, the probability density function of the generation capacity in an electricity system *P(n)* should appear similar to the function in Figure 5.6 (cf. Kling, 1998). The graph shows a binomial probability distribution for a simple system with 40 plants with equal generation capacity, an average availability of 85% and completely independent outages.

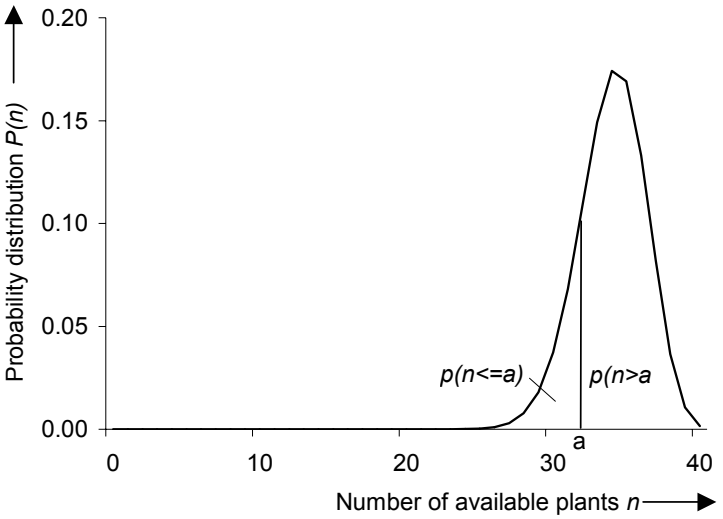


Figure 5.6: Example of a probability density function of generation capacity

The probability distribution for an actual system with plants of various sizes, with varying outage rates and related outages, such as due to cooling water restrictions, is much more difficult to establish. The basic result remains, however. The probability density function (which will be called *g(q)*, with *q* the available volume of generation capacity) starts at zero: the probability that less then 0% of installed capacity is available is zero. For all values between 0 and 100% of installed capacity *k* (MW), the probability density function has a value greater than zero. For low values of *q*, however, the probability density function is very small: the probability that only a small part of installed capacity is available is small. The function peaks close to 100%, for instance at 85%. Then it declines rapidly to zero: the probability that available generation capacity is greater than 100% of installed capacity is, by definition, zero. Thus, the general shape will be that of Figure 5.6. As it is known with certainty that the available volume of generation capacity *q* is between 0 and the total volume *k* (which is 40 in Figure 5.6):

$$\int_0^k g(q)dq = 1 \qquad (5.10)$$

The probability that the number of available plants is smaller than a is given by:

$$p(q \le a) = \int_0^a g(q)dq \qquad (5.11)$$

The shape of $g(q)$ depends upon k, upon the outage rates of the generating units that constitute k, upon the number of generating units and upon the degree to which outages are correlated. Given a fixed total system capacity, the larger the number of generating units (and thus the smaller they are), the narrower $g(q)$ will be, *ceteris paribus*. The greater the correlation of outages, the wider $g(q)$ will be. While the form of $g(q)$ is not known precisely, the mere fact that available generation capacity is stochastically distributed is sufficient to draw some conclusions.

Due to the stochastic nature of q as a function of k, there are no means to ensure the presence of a certain volume of available generation capacity q. Only installed capacity k can be controlled. Energy-not-served as a function of available capacity $E_{ns}(q)$ will be zero if *available* capacity q exceeds peak demand. By measuring the development of peak demand over the years, it may be possible to predict with a high degree of certainty which volume of available capacity q will be sufficient to avoid outages. However, energy-not-served as a function of *installed* capacity $E_{ns}(k)$ will only reach zero when k becomes infinite. More generation capacity lowers the risk of outages and the expected volume of unserved energy but will never cause them to reach zero. Consequently, the volume of unserved energy E_{ns} as a function of installed generation capacity k, given a certain peak load, is a continuously decreasing concave function which approaches zero as installed generation capacity approaches infinity:

$$E_{ns}(k) > 0 \text{ for } k \ge 0 \qquad (5.12)$$

$$E_{ns}(k)' < 0 \text{ for } k \ge 0 \qquad (5.13)$$

$$E_{ns}(k)'' > 0 \text{ for } k \ge 0 \qquad (5.14)$$

$$\lim_{q \to \infty} E_{ns}(k) = 0 \qquad (5.15)$$

Figure 5.7 shows the social cost of outages $C_o = V_{ll} \cdot E_{ns}(k)$. The cost of generation capacity $C_g(k)$ is also shown. (Similar to the assumption underlying (5.4), $C_g(k)$ is assumed to be approximated by $C_f \cdot k$, with C_f the (constant) fixed costs of a peaking unit.) The figure shows that at the optimal volume of installed generation capacity k^*, the expected cost (and therefore the loss of load expectation) of outages is above zero, and that total costs increase faster for errors in k below k^* than above k^*. The asymmetric distribution of the net social benefit of generation capacity around the optimum is a well-established result; Cazalet et al. (1978) were among the first to describe it; Billinton (1994) corroborated it

in a study of several North-American utilities.

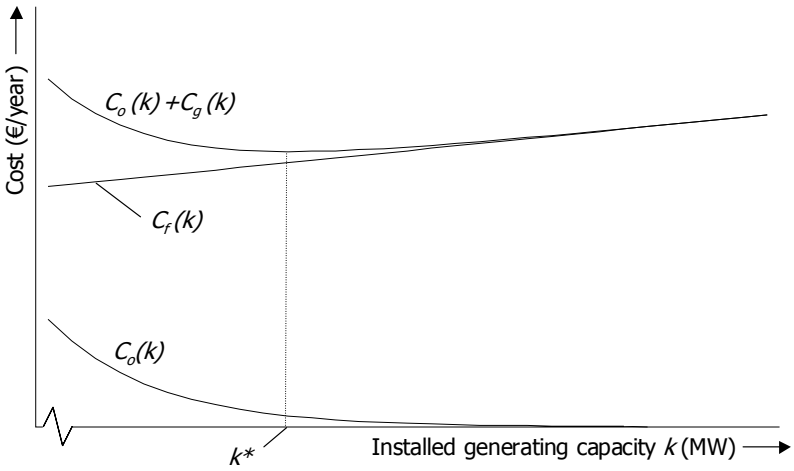


Figure 5.7: Total social cost of outages and the cost of generation capacity as a function of generation capacity.

5.4.4 Asymmetric risk

Now the question will be addressed how the cost of erring can be minimized. In particular, should we employ a strategy of conscious over or under-investment, or is there no better strategy than striving for the optimum? To investigate the nature of the optimum volume of installed generation capacity, it is sufficient to analyze the cost of outages $C_o(k)$ and the cost of generation capacity $C_g(k)$ as a function of installed capacity, as B_s was assumed to be fixed. The social optimum of generation capacity is the quantity of k for which the sum of $C_o(k)$ and $C_g(k)$ are at a minimum. $C_g(k)$ is approximated near the optimum by $C_f \cdot k$, while $C_o = V_{ll} \cdot E_{ns}(k)$. Figure 5.7 shows the shape of these functions and their sum, as well as the optimal volume of generation capacity Q^*. Because $C_o(k)$ declines with a decreasing rate, the sum of the two functions is asymmetric around the minimum.

While $C_o(k)$ is not known, some order-of-magnitude estimates can be made from recent data from the electricity crisis in California in 2000 and 2001. The outages in California totaled 30 hours, spread over six days. If the average value of lost load is on the order of 10,000 $/MWh (similar to what was calculated for the Netherlands), then the social cost of the outages was on the order of 300,000 $/MW of insufficient generation capacity. To place these figures in perspective: during 0.3% of the year, a maximum of 1000 MW, or about 2% of electricity demand, was not served (Hawkins, 2001).

Another estimate put the cost of the crisis to consumers at $ 45 billion (Weare, 2003). A majority of this cost consisted of wealth transfers and therefore does not constitute a loss

of economic efficiency (which were for a large part due to price manipulation).[19] However, even if it is assumed that only 1% was actually a loss of economic efficiency, the cost would still be $ 450 million. If it is assumed conservatively that 1000 MW was interrupted each of the 30 hours, the cost per megawatt of insufficient generation capacity is 450,000 $/MW.

These conservative cost estimates are on the order of magnitude of the fixed costs of generation capacity. This means that if a crisis of this order of magnitude occurred only once in the life span of a generator, the break-even point would have been reached, beyond which the availability of more generation capacity would not be economically efficient. Caution is required with respect to this conclusion, however: the estimates of the loss of economic efficiency due to the shortage are quite rough, while it is difficult to make more precise estimates. In addition, in a dynamically developing market, it is difficult to make the notion of an optimal volume of generation capacity operational. Given the growth of demand, additional generation capacity would be required. Finally, the estimated loss of economic efficiency is much smaller than the costs to consumers. Considering the extremely high income transfers that occurred during the crisis in California, consumers would have been better off with a larger volume of generation capacity, the cost of which would have been offset by a reduction in income transfers through the exercise of generator market power.

Even a limited shortage of generation capacity may easily cause economic losses that are on the order of the costs of generation capacity. Especially the exercise of market power may give rise to large income transfers from consumers to producers. The social costs of excess investment, on the other hand, appear limited in comparison. Shuttleworth et al. (2002) calculate that if the economically optimal reserve margin were 8% of installed capacity, and the reserve margin somehow was established at 20%, the associated social cost would be about 1.1% of the retail price of electricity.

The conclusion presents itself that the provision of electricity is characterized by a strongly asymmetric loss of welfare function, as was already suggested in Section 3.6.4 and depicted in Figure 5.7.[20] This result was first developed by Cazalet et al. (1978) and is corroborated by Billinton (1994), whose model of a sample electricity system also shows a strongly asymmetric loss of welfare function. In the presence of a highly asymmetric loss of welfare function, the likelihood of underinvestment due to the factors that were described in the previous section presents a serious risk to society, which is worth considerable cost to avoid.

Two strategies to reduce the risk of underinvestment to society can be proposed. One strategy is to 'flatten' the investment optimum by changing the dynamics of the electricity system. If demand can be made more responsive to price, a shortage would result less quickly in random rationing but first lead to the least valuable loads reducing

[19] Income transfers should not affect economic efficiency. However, an important purpose of liberalization was to lower consumer prices, in which respect they do matter. Large income transfers from consumers may prompt political intervention, as in California. See also Section 5.3.3 about regulatory uncertainty.

[20] For a similar view with respect to transmission capacity, see Hirst (2000).

their demand. This would reduce the social cost of a shortage from the average value of lost load to the value of lost load of the least valuable customers. This strategy is attractive but requires technological, institutional and behavioral changes, the potential of which is not yet certain. A second strategy is to purposely overinvest in the electricity system to a limited degree. While the over-investment would constitute a loss of welfare with respect to the social optimum, it can be considered as a social insurance against the greater risk of underinvestment.

Small as the cost of limited excess capacity may be in a perfectly competitive electricity market, in less-than-perfectly competitive market there may be additional benefits to consumers that may even outweigh the costs. Excess generation capacity would reduce the prevalence of opportunities to exercise market power through capacity withholding. (Section 5.6 will discuss market power.) Considering the oligopolistic nature of many electricity markets,[21] the decrease of income transfers from consumers to producers as a result of a reduction of market power may offset the costs of the extra generation capacity. This is an additional argument for erring on the side of excess generation capacity. Chapter 6 discusses market design alternatives to reduce the risk of underinvestment, some of which also provide better incentives for demand-side management.

5.4.5 The perspective of generating companies

Before liberalization, the vertically integrated utilities developed a mix of base, medium and peak load units and retained old units as a reserve. Their primary concern was with system reliability, as they could usually pass the costs along to the consumers. In a competitive market, the motive for investment is profit. Generating companies no longer are responsible for system reliability. They cannot be, as they only serve part of the market. The question is whether in a competitive market investment decisions that are optimal from the perspective of generating companies lead to an outcome that is desirable from the perspective of society. In this section the argument will be made that generating companies, like consumers, have an asymmetric loss of welfare function with respect to the investment optimum, but one that is reversed.

An increase in peak demand is not necessarily met with an investment in peaking capacity. Large generating companies with a portfolio of generating units have a choice between expanding their generation capacity by investing in base or medium load, or by investing in a peaking unit with low fixed costs and high variable costs. The latter case, investment in a peaking unit, is the one that was considered implicitly in the analysis of this chapter. Calculating the rate of return is simple in principle: the fixed costs need to be recovered during the hours that the unit operates, as was shown in Example 5.1.

When a generating company invests in a plant that is higher in the merit order, calculating the rate of return is slightly more complex, as the new plant changes the merit order of the company's generators. Therefore the impact of the investment upon the generation portfolio needs to be considered. Again, the investment cost consists of the

[21] See footnote 26 on page 99.

fixed costs of the new plant but the benefits consist of two parts. First, the new plant brings about a reduction of operating costs along the company's entire supply curve upwards from the new unit. Second, the generating company's total capacity is expanded, so it can sell a higher volume during price spikes. The degree to which the reduction in operating costs translates into higher returns depends upon the shape of the general supply curve and the degree of competitiveness of the market. The price may drop if the capacity addition leads to a different the marginal generator.

Figure 5.8 shows the effects upon the generating company's supply curve of investment in peaking capacity on the left and investment in medium load generation capacity on the right. The bold line indicates the new generator: investment in a peaking plant (left) adds a new plant with high variable costs at the end of the supply curve, whereas investment in a medium load plant adds new capacity in the middle of the company's supply curve. Investment in peaking capacity therefore simply extends the supply curve, whereas investment in medium load capacity shifts part of the supply curve to the right, as a result of which operating costs are lowered for a range of output. If the same volume of new generation capacity is considered in both options, the trade-off is between the lower fixed costs of a peaking unit versus the benefits of lower operating costs of a medium load plant.

In the case of investment in a medium load plant, the generating company may decide to retire his most expensive peak load plant. The question of whether the plant will be retired depends upon its expected price spike revenues versus its fixed costs (such as capital, a crew, maintenance and fuel contracts). If the plant is retired, the new medium load plant was simply a replacement investment, which did not affect overall capacity.

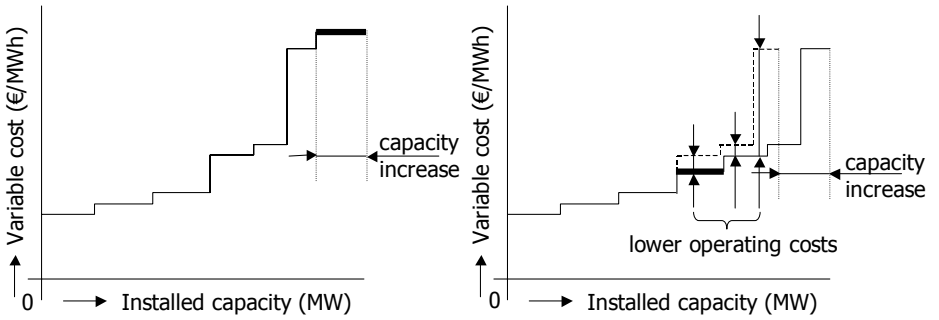


Figure 5.8: Investment in peaking capacity versus medium load capacity

The previous sections argued that investment in peaking capacity is not likely to be socially optimal, in large part because it is too risky. The question is whether investment in base and medium load can compensate. When new generation technology offers substantial cost savings, it may be an attractive investment due to the reduction in operating costs. Electricity generating technology is mature, however, so the variable costs of a new plant are only slightly lower than those of existing plant. This means that an important part of the investment must be recovered through the increase in turnover which is made possible by the new plant (unless it is a replacement investment and the

old plant is decommissioned). As a result, calculating the profitability of the investment in a medium load plant becomes similar to that for a peak load plant, with the same risks and uncertainties. (See Example 5.1 on page 83.)

A new market entrant faces a different scenario. As he does not have a portfolio of generators, his only concern is that the average price will exceed the total costs of his new plant. The new plant will change the aggregate supply curve of the market (the combined supply curves of all generating companies), and the marginal unit will see its expected operating time decrease. The owner of this unit will need to re-evaluate whether he expects the fixed costs of this unit to be recovered during price spikes. Again, this will depend upon the factors that were described in Example 5.1.

In any case, the total volume of generation capacity depends upon the profitability of the marginal generator. The owner of the plant will weigh the fixed costs against the projected height, frequency and duration of price spikes, which are all highly uncertain, as was seen above. As a result, for the sake of our analysis, only the profitability of the marginal plant needs to be considered.[22] What is the impact of these uncertainties upon investment?

Traditionally, investment projects are evaluated with the Net Present Value (NPV) method (cf. Ross et al.). In this method, uncertainty about future returns is reason to use a higher discount rate. The significant sources of uncertainty in current European markets that were discussed in Section 5.3 would be cause for a high discount rate, which would reduce the present value of future returns. All else being equal, this would reduce the equilibrium volume of generation capacity.

The NPV method is criticized, however, for neglecting the opportunity cost of investing when projects are irreversible (Dixit and Pindyck, 1994). The NPV method only provides an indication whether an investment is worthwhile or not but does not consider the value of delaying the project in the presence of uncertainty about future revenues. Postponing an investment decision may provide a benefit of better information regarding the profitability of the project. Therefore, if an investment is irreversible, there is a cost associated with forgoing the opportunity to wait for new information. This opportunity cost must be weighed against the costs of waiting (such as a loss of revenues or allowing a competitor to take the initiative), according to Dixit and Pindyck (1994). An extra reason to postpone investment in electricity markets that are being, or recently have been, restructured is that the restructuring process itself creates significant uncertainty, which may be expected to decrease as the market matures. Restructuring of related markets, such the natural gas market in Europe and the new market for tradable CO_2 emissions permits, further increases the premium on waiting.

Power plants are a perfect example of irreversible investments and electricity markets are characterized by significant uncertainty. Therefore the real options theory, as outlined by Dixit and Pindyck (1994) best describes their investment decisions. A key result is that in

[22] The marginal plant does not need to be constructed for the purpose of providing super peaks, but that it may be an old, written-off plant which is kept stand-by. This reduces the fixed costs of this plant, but does not change the nature of the argument.

the presence of uncertainty, the real value of an investment project may be significantly lower than the net present value. Consequently, firms require above-normal rates of return to justify investment. Similarly, existing generators are not retired as soon as the price drops below the long-run marginal cost because retiring the plant forecloses any opportunity to make profits if prices rise again. Only if prices fall dramatically low for a prolonged period will this lead to the closure of plant. As a result, the impact of the electricity price upon investment is characterized by hysteresis: there is a bandwidth around the long-run marginal cost around which prices neither trigger investment nor disinvestment (Dixit and Pindyck, 1994).

This means that the development of the industry is path dependent. When prices rise well above the long-run marginal cost, they attract investment that does not disappear when prices return within the bandwidth around the long-run marginal cost. Thus, the development of generation capacity lags behind the development of price. This suggests a tendency towards the development of investment cycles. The next section will explore this phenomenon further.

According to the real option theory, the option value of waiting must be weighed against its cost, for instance the risk that a rival will invest first and thereby reduce the expected returns. In an oligopolistic market structure, the latter risk may be smaller than in a perfectly competitive market. In the presence of barriers to new market entrants, the incumbent firms may deploy a strategy of waiting until the uncertainty decreases. Therefore it is to be expected that this tendency is stronger in an oligopolistic market structure (see Section 5.6.).[23]

5.4.6 Summary

Determining the socially optimal volume of generation capacity requires data that are difficult, if not impossible, to obtain with sufficient accuracy. The same data would be required for generating companies to make profit-maximizing investment decisions. Given uncertainty about the optimal volume of generation capacity, the prudent policy for society is to err on the side of extra generation capacity, as the social costs of investing too little appear at least an order of magnitude higher than the costs of investing an equivalent amount in excess of the theoretical optimum. Generating companies, however, have an incentive to delay investment, given the many uncertainties that characterize current electricity markets. It may be concluded that in the presence of uncertainty with respect to future electricity demand, the public and the private interests do not coincide. Consequently, the design of electricity markets should not only focus upon providing generating companies with theoretically optimal incentives but also with the possibility that the optimal volume of generation capacity is not obtained.

[23] See also Neuhoff and De Vries (2004) for an analysis of the impact of risk aversion among generating companies and consumers. This article demonstrates in a different way that risk-averse generating companies would invest less than risk-neutral ones, while risk-averse consumers would prefer a higher volume of generation capacity.

5.5 Long-term market dynamics

The previous sections argued that energy-only markets are not likely to produce an optimal level of investment in generation capacity. Shortages may lead to extreme price spikes, which should trigger investment. A prolonged period with price spikes could lead to an overreaction by investors, which would lead to an investment cycle. The lag in investment behavior which the real options theory predicts could exacerbate the tendency towards investment cycles (Dixit and Pindyck, 1994; see also Section 5.4.5).

This section analyzes the possibility of the development of an investment cycle. It argues that even if average generator revenues would be sufficient to cover the costs of the optimal volume of generation capacity, the high volatility of electricity prices combined with imperfect information could lead to investment cycles. These would lead to a higher incidence of shortages than would be optimal. Factors disturbing the optimal investment equilibrium, such as discussed in the previous section, would make things worse. In the first part of this section a closer look will be taken at the possibility of investment cycles. The second part describes why the obvious solution, long-term contracts, will not solve the issue of generation adequacy.

5.5.1 Investment cycles

The electricity shortages in California in 2000 and 2001 appear to have been caused, at least in part, by lagging investment in generation capacity (see Chapter 4). While there is strong evidence that illegal withholding of generation capacity played an important role in the development of the crisis, and even may have been the immediate cause of a substantial part of the power shortages, the opportunity to exercise market power developed as a consequence of the narrow capacity margins during the crisis.

A year before the beginning of the crisis in California, Ford (1999) published a paper based upon a computer simulation, in which he showed that investment in electricity generation facilities is inherently unstable in a system with rules such as in California. His explanation is that investment is not aimed at dampening business cycles (which it would do if the right amount of new capacity became available at the right time) but at making a profit. Ford assumes investors are risk-averse and tend to wait until they are reasonably certain that they can make a profit. In his model, they also tend to overreact, in part because they may not know their competitors' plans. Therefore Ford (1999) considers the interaction between a mandatory power pool and investors inherently unstable. He notes, however, that capacity payments may dampen the investment cycle. The issue of how to prevent shortages in electricity markets, structural or periodic, will be discussed in Chapter 6.

A fundamental cause of investment cycles is the time lag between an investor's decision to build new generation capacity and the moment it becomes available for power production. Ford (1999) showed that the presence of this delay leads to investment cycles even if external factors such as the costs of new plants, fuel costs and demand develop in a relatively predictable manner. When the investment environment is less stable, a long investment lead time increases the uncertainty with regard to future demand and

operating costs. This will exacerbate the cyclical behavior, as the increased uncertainty causes investors to wait longer.

Visudhiphan et al. (2001) contend that a lag time need not cause investment cycles, as long as the investors are able to anticipate market developments. However, as was seen above, sufficient information about future supply and demand generally is lacking. In their simulation, Visudhiphan et al. also find that backward looking investment, that is, investment based upon recent experience in the market, will lead to investment cycles. Stoft (2002) arrives at the same conclusion. He notes that the distribution of price spikes may be such that investors would need to have a time horizon of several decades to determine the real average revenues from price spikes. If they use a shorter time horizon, they are bound to overestimate or underestimate their expected revenues.

The Appendix (page 301) presents a simple dynamic model with which dynamic investment behavior is simulated. In this model, investment cycles result from a combination of imperfect foresight by investors (they do not anticipate the future growth of demand accurately) and a delay between the decision to invest in new generation capacity and the moment this new capacity becomes available for electricity production. Generating companies are modeled to invest mainly in reaction to high prices, as a result of which the new capacity arrives too late. The purpose of this model is not to prove the existence of investment cycles but to study the potential effect of the capacity mechanisms that will be discussed in the next chapter. The model is introduced here as an example of what an investment cycle might look like.

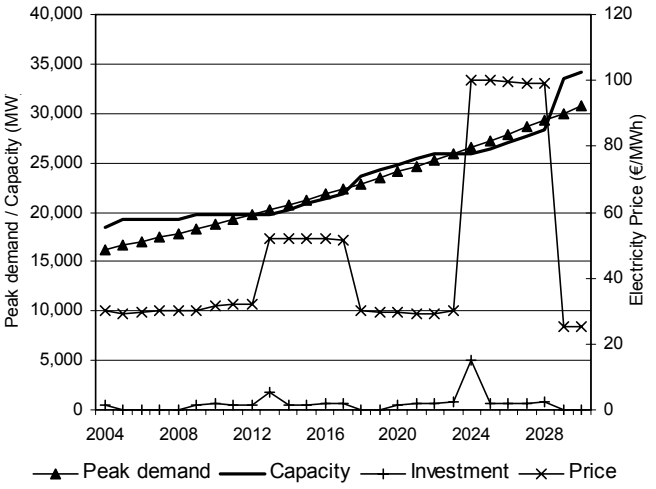


Figure 5.9: Investment cycle in the presence of a fixed growth rate of demand, which is under estimated by investors

Figure 5.9 shows the model results for a system with a peak demand of 16,200 MW in 2004 and an annual demand growth rate of 2.5%. The X-axis shows the years from 2004

through 2030. The left Y-axis shows capacity, both total available generation capacity and the volume of new investment, and peak demand. The right-hand Y-axis shows the annual average electricity price. See for a detailed description of the model the Appendix.

Investors are modeled to anticipate a growth rate of 1.5% per year, reflecting a paucity of information and/or risk aversion. Price spikes induce additional investment. Overcapacity at the start of the modeled period means that the first number of years there is ample generation capacity. A first shortage occurs between 2013 and 2017. During this period, the volume of interruptible contracts (modeled to be 500 MW) is just sufficient to avoid outages. The high price of these contracts (modeled to be 2500 €/MWh) leads to a succession of price spikes, which induce just sufficient investment to meet demand during the next several years. Towards the end of the modeled period, in 2024, another shortage starts that does lead to outages. The price spikes now go up to the average value of lost load, which is assumed to be 8,600 €/MWh (based upon Bijvoet et al., 2003).

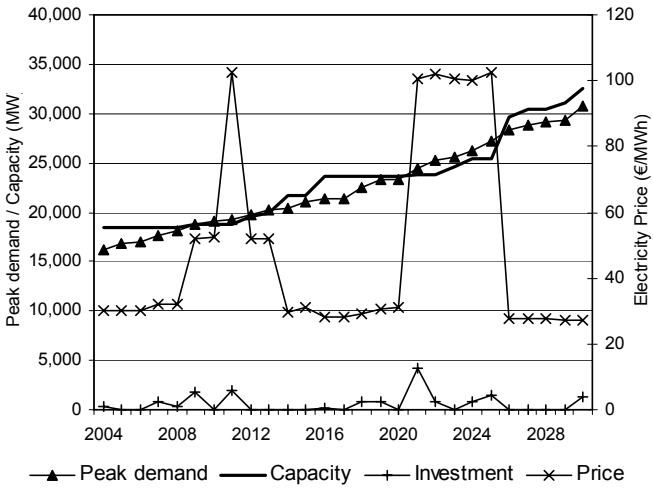


Figure 5.10: Investment cycle, demand growth fluctuates randomly around 2.5%, demand forecasts are an extrapolation of historical trends

The model shows a relatively smooth development of the total volume of generation capacity, which is due to the fact that a constant growth rate was used. If the growth rate fluctuates randomly, as in Figure 5.10, a period of shortages may occur much sooner. In this case, the investors' forecast of the demand for generation capacity was based upon an extrapolation of the growth rate of the last five years. No bias was included in the investment behavior, so investment takes place in a risk-neutral manner, but based upon historical data. As a result, periods of low growth lead to an under-estimation of the future need for capacity. A higher growth rate of demand, or shocks, such as a reduction of the availability of electricity for imports or a decision to phase out a certain generation technology, may cause shortages to arise sooner.

The argument of investment cycles is fundamentally different from the analysis presented in Sections 5.3 and 5.4. These sections presented a static analysis, rooted in neo-classical economics. There, the assumption was that a market equilibrium would develop; the analysis was focused on the question how to ensure that that equilibrium was socially optimal. The argument that investment cycles may develop is a different one. Even in the presence of theoretically perfect incentives, the market outcome may be far from optimal, if it involves significant oscillations around the investment equilibrium. The possible causes of investment cycles are those that were described in Section 5.3 and 5.4. The time lag of new generation facilities, insufficient information about the supply and demand functions, and risk aversion are causes of the development of investment cycles. In addition, the public good character of reserve capacity is a factor, in the sense that marginal peaking units are undervalued during off-peak times, when their income is zero, and perhaps overvalued during price peaks, depending on the maximum price. As these factors are difficult to remove in energy-only markets, the risk of investment cycles should be taken seriously. The potential social damage may be considerably higher than a static analysis would suggest, as the cycles could create large deviations from the socially optimal level of generation capacity.

5.5.2 The role of long-term contracts

A classic solution to provide stability to the market consists of long-term contacts between producers and consumers. They could greatly reduce investment risk by providing financial stability and by making the demand for capacity explicit. Although currently only base and medium load capacity is sold in long-term contracts, many of the potential causes of market failure could be removed if power from peaking units were also contracted on a long-term basis. To cope with the uncertainty regarding the actual demand for peak power, long-term contracts for peaking capacity presumably would take the form of call options, consisting of a fixed payment that gives the buyer the right to purchase electricity at a specified price. Such contracts would remove much of the price volatility, which is a risk for generators and consumers alike. If both sides benefit, why are peaking units not covered by long-term contracts in practice?

Public good

There are several problems with long-term contracts. The first and main one is the fact that reserve capacity is a public good if there is a possibility of service interruptions due to a shortage of generation capacity and if it is not easily possible to interrupt electricity service to individual consumers based upon their willingness to pay. Under these conditions the reliability is the same for all consumers who use the same network. (See Section 5.2.3.) Consumers who engage in long-term contracts in order to enhance their own reliability of service, improve the reliability of service for everyone else in the system as much as for themselves. Consumers have an incentive to free-ride upon each other, as paying more for reliability would be akin to paying a voluntary tax. This argument holds whether the generating companies deliver their electricity directly to consumers or via retail companies. In the latter case, retail companies have a disincentive to purchase long-term peak load contracts from generators (Neuhoff and De Vries, 2004). If these contracts are to provide an adequate signal to generators to install sufficient

generation capacity, they must pay the generators the average cost of peaking capacity. During periods in which the peaking capacity is not used (which is actually most of the time), the spot market price will be below the cost of these contracts, as competitive spot market prices reflect the marginal cost of production. Retail companies that hold long term peak load contracts will therefore have higher costs, as they contribute to the capital cost of the peaking unit, and not do well on the market.

The exception is when retail companies have a regional monopoly over the delivery of electricity. In this case, their generation contract portfolio determines the reliability of service in their region. This is an argument for preserving the consumer franchise, as it provides a simple solution to the question of generation adequacy (Newbery, 2002a). However, it would mean that liberalization would be limited to large consumers, while the captive consumers would pay for the reliability of the entire system.

Slow learning curve

Even if generating companies would receive the appropriate long-term demand signals, the physical inertia of the electricity sector would present an obstacle. The long time it takes to develop new capacity and the long life cycle of generating plant makes it difficult to reach an efficient equilibrium (cf. Vázquez et al., 2000). If investment signals depend upon consumers entering into long-term contracts to hedge their risk of supply interruptions, consumers need to have the opportunity to learn which contracts are attractive to them. As they would mainly learn through trial and error, this would require repeated periods of shortages and high prices. The physical inertia of the electricity sector and the close relationship between demand growth and the general economy cause the business cycle of the electricity sector to be long, probably on the order of a decade.[24] As a consequence, consumers have few opportunities to learn how to find attractive contracts that hedge their risk of supply interruptions.

It is likely that a period of shortages leads to an adjustment of the regulation of the market, so the learning curve would start over. The result would be that consumers would never learn to cover all of their future demand with long-term contracts, so electricity shortages would reoccur time and again and the market would never reach an equilibrium. So even if end consumers or their suppliers would have a proper incentive to enter into long-term contracts for peaking capacity, it would take unacceptably long before they would know what their actual (long-term) needs are and how to negotiate these contracts.

Long business cycle

A final practical problem with long-term contracts is that the duration of such contracts in electricity markets is generally too short to dampen the business cycle. In principle, in the presence of sufficient information, long-term contracts should reflect the average expected spot market price. Would they be any different, either suppliers or consumers would consider them unattractive. However, this requires both investors and consumers to be able to estimate the average spot market price over the entire business cycle of

[24] In the model of the Appendix, the frequency of investment cycles is about once every 12 years.

generators. Instead, if prices are low for a number of years in a row, market parties may assume that these prices will remain low for the foreseeable future and base their contracts upon them. As a result, a period of low prices will lead to under-investment. Once demand growth has absorbed all excess capacity, a period of scarcity and high prices follows. According to the same line of reasoning, this in turn may lead to over-investment, as a result of which prices will decrease again. The time horizon of long-term contracts is unfortunately too short to dampen the business cycle, so they fail to ensure the reliability of service.

Even if generators would be willing to consider long enough contracts to average the swings of the business cycle, the risk to consumers of such contracts would probably be too large. In the course of a decade or longer, the fuel markets are likely to change, generation technology may change and the uncertainty about the development of demand is large. This is the *Catch-22* of long-term contracts: a short time horizon does not isolate the contracts sufficiently from the business cycle, while a long time horizon carries too much risk.

Conclusion

Investors lack the incentive and the time horizon to engage in sufficient long-term contracts. Consumers do not know the value of long-term contracts and will tend to buy from retail companies without long-term contracts when prices are low. Therefore it may be concluded that while long-term contracts may cover a significant portion of generation in a mature market, they cannot be expected to cover peak load capacity. Especially during a period of excess capacity and low prices, a shortage of long-term contracts for peaking capacity appears to be likely. Consequently, it is to be expected that this period of excess capacity is followed by a power shortage, and that an investment cycle develops.

5.6 Market power

5.6.1 Short term: withholding during a shortage

A different weakness of relying upon periodic price spikes to signal the need for investment is that, as long as demand is relatively price-inelastic, these price spikes may provide perverse incentives to generating companies. When high price peaks occur, there is a strong incentive to withhold generation capacity from the market in order to further increase the price. This was a significant factor in the crisis in California, which contributed both to the extreme height of the electricity prices and to the service interruptions (Joskow and Kahn, 2002). The reason is that when the capacity margin is slim, or when acute shortages already exist, the low price-elasticity of demand means that a small reduction in the supply of electricity may lead to steep price increases. In such a situation, generating companies are able to increase their revenues by keeping some generation capacity off the market, which results in a price rise which more than off-sets the lost volume of sold electricity.

Stoft (2002) points out that if there is no price cap if it is very high, for instance equal to the average value of lost load, the increase in profits from withholding can be so high that it becomes attractive even for small generators who would have to withhold a majority of their generation capacity. As a result, many generating companies, not just the large ones, have market power during a period of scarcity. The increases in profit that result from withholding may be large, while it is difficult to take juridical steps against this behavior – if it is illegal at all. One would have to prove which outages were illegal, rather than forced, for each time unit during which withholding is suspected. From the point of view of generating companies, the only disadvantage of this strategy, besides the possibility of being caught for abuse of market power, is that withholding electricity during a period of scarcity may cause such a political crisis that it prompts a complete overhaul of the market design, as it did in California.

Long-term contracts limit the incentive to withhold capacity.[25] If 90% of the market is covered by long-term contracts, generators' potential gains from capacity withholding are reduced by a factor of ten. Therefore the development of long-term contracts is highly desirable, even if they cannot be counted upon to provide enough incentive to invest. A vulnerability is, however, that their duration may be too short in comparison to the prolonged period of scarcity which may result from an investment cycle. As it may take several years to construct new generation capacity, such a period may last that long. Many long-term contracts last a year or less, so a significant portion of them may expire during an episode of scarcity. In this case, capacity withholding becomes increasingly attractive again, not only to increase the short-term gains in the spot market but also because the spot price serves as a reference point for new long-term contracts.

A larger volume of generation capacity limits the average amount of time that the system is short of capacity, and therefore also the possibility for market power in the short-term market. This adds to the asymmetry around the investment optimum that was observed in Section 5.4. As the abuse of market power increases the cost of a shortage to consumers, it becomes even more attractive to avoid shortages in a market that is less than perfectly competitive.

5.6.2 Long term: strategic investment behavior

European electricity markets exhibit strong oligopolistic characteristics (AER, 2003).[26] If there is no strong price competition, an oligopoly of large generators may choose a more stable, long term strategy, rather than to create or exacerbate a power crisis through the withholding of generation capacity. If generation companies are able to keep prices above the competitive level during normal market conditions, they may even opt to overinvest in order to discourage new entry. Because investment in generation capacity is irreversible and the large sunk costs of power plants make exit from the market difficult, the presence of generation capacity serves as a credible threat to newcomers that the

[25] Cf. Allaz and Vila, 1993, who show that forward contracts improve competitiveness

[26] For example, the French, Belgian, Portuguese, Italian, Greek, Danish and Irish markets are dominated by one or two generators, while only three or four producers serve two-thirds or more of the markets in Germany, Austria, Sweden, the Netherlands and Spain (EU energy markets, 2002).

incumbents will continue to stay in the market (Spence, 1977). Another reason for an oligopoly to invest more than would appear to be economic, would be to enhance reliability. A stable oligopoly may place an extra value upon reliability because service interruptions would attract undesired (political) attention. However, this hypothesis may not always hold, especially in the presence of uncertainty about future demand (Tirole, 1988).

Von der Fehr and Harbord (1997) note several effects in oligopolistic markets that counteract each other, so they consider it unclear whether an oligopoly generally will invest too little or too much. Profit maximization would lead to underinvestment, as would the effect that investment in the lower or middle reaches of the supply function (in base or medium load capacity) would 'flatten' the supply curve, thereby lowering expected revenues for all generation capacity. These effects may be countered, according to Von der Fehr and Harbord, if short-term market power causes spot prices to rise above the competitive level, which would tend to lead to overinvestment. Note that these effects need not occur simultaneously. They could reinforce the tendency towards investment cycles which already was observed in competitive markets. Sufficient market power to increase spot prices may only exist during a shortage, whereas during periods of excess capacity the presence of market power may enhance the tendency not to invest.

It may be concluded that an oligopolistic market structure may counter the tendency to underinvest that was found in the previous analysis. In this situation, a regulator may find himself facing the paradox that stimulating competitive behavior, for instance through the application of competition law, would reduce reliability. Allowing an oligopoly to exist is hardly an attractive option, however, because it would undo many of the efficiency gains from liberalization. In addition, it would lead to prices above the competitive level, which would attract imports, increasing dependency upon other systems for system reliability (Newbery, 2002a; see also Figure 4.6 on page 54). This presents policy makers with the dilemma whether to allow the oligopoly to exist, or whether try to increase the competitiveness of the market, which could reduce investment in generation capacity.

Incumbent generating companies may be assisted in their strategy by the presence of other barriers to entry. It may be difficult for new market entrants to obtain sites for power plants, especially in densely populated areas. Incumbents may be able to re-use the sites of decommissioned plant, where they often already have a connection to the high tension grid and infrastructures for fuel and cooling water, and where they will probably face less difficulty in obtaining the necessary permits.

If the balancing market is not efficient, this also poses an obstacle to new market entrants: generating companies with more generating units are better able to handle their imbalances themselves than a small generating company with few units. Therefore the cost of unscheduled outages will be smaller for large generating companies. Vertical integration may also pose a barrier to entry, as it reduces the liquidity of the market. Newcomers may want to secure their output with long-term contracts to avoid the risk of not having customers. These factors, if present, would support a strategy of deterring entry with overcapacity.

5.7 Technological changes in the electricity sector

If only one of the characteristics of the electricity sector that were presented in Section 5.2 changes, the dynamics of the market will be fundamentally different. Section 2.4.3 outlined some possible technical developments. This section briefly discusses how they may impact the issue of generation adequacy.

Electricity storage

Currently, different techniques to store electricity are being developed, some of which are approaching the point where they may be commercially viable (see for instance Regenesys, 2003). A storage device that could store electricity at a cost smaller than the price difference between daily peak and base load prices would significantly impact market dynamics. Peak load units would no longer be the commercially most attractive means of meeting peaks in demand; rather, electricity generated in base or medium load plants would be stored and released during demand peaks. Storage would lower the peak demand for generation capacity, as a result of which price spikes also would become lower. Investment risk would decrease, so the tendency towards investment cycles would also be dampened.

Real-time metering

Another characteristic of the electricity sector that stands in the way of an efficient market, the low observed elasticity of demand, may also change over time. It is often assumed that the actual demand elasticity is higher than observed but that the institutional arrangement of electricity markets artificially creates a nearly inelastic demand function. Regulated end user tariffs for captive consumers completely take away any incentive to adjust consumption to the price of electricity but also in fully liberalized markets demand often appears to be highly inelastic. Most consumers do not know the real-time price of electricity. Moreover, bills often only present an average cost per kilowatt-hour, so even *ex post* information about the cost of using electricity at specific times is absent.

Different experiments have shown that at least certain categories of customers are quite willing and able to adjust their demand to electricity prices (cf. Roberts and Formby, 2001; Sæle and Grønli, 2001). During the capacity shortages in California in early 2001, consumers showed a considerable demand elasticity by reducing peak demand by up to 12% on a voluntary basis, once the severity of the crisis had become clear (Coleman, 2001). A more price-elastic demand would probably not lead to a much lower overall demand for electricity but consumers would shift some of their electricity use to periods with lower prices. This would result in a flattening of the demand peaks, which has the same impact upon generation as the use of storage devices. Improving demand price-elasticity is a matter of consumer education and, most importantly, investing in the necessary communications infrastructure to provide consumers with the necessary information. In addition, consumers may need to invest in equipment that can help them program their loads, for instance timers or devices that switch off loads if the electricity prices exceed a specified level. Implementing these arrangements on a large scale would take considerable time and investment while their full potential is not certain.

Distributed generation

In the past, efficiency improvements led to ever larger power plants. Combined-cycle gas turbines have reversed the trend, as they can achieve a high efficiency in small units. Fuel cells, if they break through, carry the same promise. This has lead to speculation that in the future, electricity generation will be distributed in nature, with many small power plants near consumers, rather than a few large ones far away.

In the analysis of this chapter, the scale of power plants plays a limited role. Nevertheless, a shift towards distributed generation could have significant effects. First, it is easier to estimate the availability of generation capacity when there are many small units, as the deviation from the average will be smaller. (In the terms of Figure 5.6, *g(q)* would be 'narrower'.) A second effect could be that the lead time for new capacity would be shortened substantially, for instance if the generating units would be serially produced and could be delivered from stock. This would reduce the tendency towards investment cycles. A shorter lead time and smaller units could also facilitate market entry. This could reduce market power but that would also depend upon other factors: there may be economies of scale in bundling the output of many small generators on the market, for instance to manage imbalance issues.

It may be concluded that changes in technology may reduce the volatility of electricity prices, which would reduce investment risk and therefore also reduce the tendency of the sector to develop investment cycles.

5.8 Trade between electricity systems

A different aspect of the issue is how to ensure generating adequacy in the presence of significant volumes of trade between systems. In theory, trade between liberalized electricity systems should not change the fundamental nature of the market dynamics. If the connected systems are liberalized in similar ways, trade between them only represents a scale increase. The scale of the system does not change the issue of generation adequacy, as it was addressed in this chapter. A benefit of a larger interconnected system is better operational stability, as the relative impact of individual generators and capacity additions becomes smaller.

In practice, however, interconnected electricity systems often have quite different market rules, and the rules for using interconnectors are different from the regular transmission access rules within the systems. Therefore the interconnected markets are not fully merged. This has repercussions upon the generation adequacy in the different markets.

In the case of California, part of the problem was that investment in generation was not only lagging in California itself but also in neighboring states. There, however, it did not lead to a shortage but only to a reduction of the supply margin. When the weather suddenly caused a shortage, these states were able to use their own generation resources for their own demand first, selling to California only any excess electricity. As a result, California, the importing state, bore the full brunt of a crisis the roots of which actually were spread among a number of states.

In Western Europe, a similar scenario is possible. Article 24 of the Directive allows member states 'in the event of a sudden crisis' to take unspecified 'safeguard measures' (Directive 2003/54/EC). This can be interpreted as giving member states the right to close down interconnectors temporarily in an emergency.[27] While there may be technical reasons for this, this means that in the case of a crisis, the European internal market may fall apart into several unconnected markets. This complicates the analysis of generation adequacy in a specific system. On the one hand, questions such as the optimal volume of generation capacity and how investment should be stimulated to reach this optimum must be considered in the day-to-day reality of a large, interconnected system with trade. On the other hand, the risk of temporary reductions of interconnector capacity raise the question of whether reliability can only be maintained by having sufficient generation capacity to be self-reliant.

A second complication arises when interconnectors are congested. The volume of available interconnector capacity depends upon the load flow and may therefore vary to a degree. Maintenance may further affect the availability of transmission capacity. Finally, the limited periods for which import capacity typically is auctioned prevents importing parties from engaging in long-term contracts for generation capacity.

5.9 Policy choices

Despite the concerns about the stability of energy-only markets, it is not certain that they will indeed fail to provide a socially acceptable volume of generation capacity. There is neither sufficient empirical information nor such a highly developed understanding of the dynamics of the electricity system that the conclusion may be drawn with certainty that policy intervention is warranted. At the same time, society does not wish to accumulate much experience with failure of the electricity generation market.

An intuitively attractive policy is to monitor the development of the market and to intervene when investment appears to be insufficient to guarantee future generation adequacy. The problem is the long lag between implementation of a policy to stimulate investment and the availability of more generation capacity to the market. This lag time may be a number of years: for instance a year to implement the policy, a year or more for the market to analyze its effects, plus several years to obtain the permits and build new generation capacity. A policy of monitoring would require being able to forecast the development of the margin between generation capacity and demand that far into the future. It is unlikely that a monitoring system can provide this information (Van Werven, 2003), which means that a wait-and-see policy cannot fully remove the risk of underinvestment. This leaves policy makers with a dilemma:

> *Policy choice 5.1: Should a mechanism be implemented to secure generation adequacy before there is empirical evidence of the tendency to underinvest, or should society wait and see how the market develops, which implies risking a*

[27] A question is to which degree system operators are physically able and politically willing to interrupt exports.

period of scarcity?

In the judgment of the author, the precautionary principle applies here: the fact that the magnitude of the social costs of underinvestment likely are much greater than the costs of overinvestment is a reason to intervene preemptively. The next chapter will present a decision framework and options for adjusting the market design.

In the presence of a stable oligopoly of generating companies, the nature of this question changes somewhat. If the oligopoly has a strategy of providing sufficient (or excess) generation capacity, for instance in order to deter new market entrants, policy intervention for the sake of generation adequacy may not be necessary. However, it is uncertain whether the oligopoly will (be able to) maintain its strategy. The following policy choice results:

> *Policy choice 5.2: Should a capacity mechanism be implemented in an oligopolistic generation market, even though it may have a strategy of providing excess generation capacity?*

Finally, Section 5.8 showed that importing systems are confronted with another question, if they cannot be fully certain of the future availability of the imports:

> *Policy choice 5.3: If the future availability of imports is not as certain as domestic generation capacity, to which degree should these imports be considered as contributing to generation adequacy?*

The evaluation of policy options in the next chapter will consider how a certain volume of domestic generation capacity can be obtained in the presence of significant exchanges with systems with different market rules.

5.10 Conclusions

The theory that perfectly competitive energy-only markets can provide an optimal amount of generation capacity in an equilibrium situation has a number of weaknesses. First, the low price-elasticity of demand and the inability to store electricity cause electricity prices in most markets to be highly volatile, as a result of which investment risk is substantial. This makes the equilibrium volume of generation capacity vulnerable to distortion factors.

Second, several factors further increase investment risk, such as regulatory uncertainty and insufficient information regarding future supply and demand conditions. These factors make it difficult to identify the optimal volume of generation capacity, both from the perspective of the generating companies and from the perspective of society.

Third, it was argued that given the significant uncertainty about the optimal volume of generation capacity, private and social investment equilibria do not coincide. Private investors prefer to err on the side of less generation capacity, while the prudent strategy for society is to over-invest. Not only is the cost of excess generation capacity limited, a

higher volume of available generation capacity also reduces market power, *ceteris paribus*. This further increases the incentive for consumers to err on the side of excess generation capacity.

For these reasons, and due to the long lead time for new generation capacity, there is a substantial risk that investment cycles will develop, in which periods of high prices and shortages well in excess of the optimal duration of outages are followed by periods of excess capacity and prices below the cost price of electricity.

A different type of weakness of energy-only markets is that price spikes may be manipulated, as occurred in California. Whether this happens, depends upon the degree to which generator output is exposed to spot prices. Long-term contracts reduce incentives for withholding capacity, but they may expire during an extended period of shortages.

The above arguments give reason to implement measures for securing a sufficient volume of generation capacity. Based upon the analysis in this chapter, policy intervention should meet the following requirements:

- it should stabilize the volume of generation capacity,
- it should be compatible with imports and exports, if applicable,
- it should be robust against the abuse of market power, and
- it should improve demand price-elasticity.

6 Capacity mechanisms

Several adjustments to the design of wholesale electricity markets, which will be called capacity mechanisms, have been tried in practice or proposed in the literature to improve generation adequacy. This chapter describes the principles of capacity payments, strategic reserve, operating reserves pricing, capacity requirements, reliability contracts and capacity subscriptions. Chapter 7 will provide an evaluation of these mechanisms.

6.1 Introduction

A number of liberalized electricity systems have taken measures – capacity mechanisms – to maintain sufficient generation capacity in the long term. Capacity mechanisms are sometimes considered to be transitional measures used while an electricity market is maturing, providing stability and a safety net during the sometimes turbulent transition to a liberalized market system. The analysis in Chapter 5 suggests, however, that the need for capacity mechanisms may be more permanent – at least until demand is significantly more price-elastic. An exception is when competition is not extended to the retail market (Newbery, 2002a). When at least part of the consumers are captive, the free-rider problem (Section 5.5.2) is solved. Then retail companies can cover their full projected demand with long-term contracts. This option will not be discussed further because the EU, which is the focus of this study, requires full market opening (Directive 2003/54/EC).

In this chapter, the following capacity mechanisms will be evaluated:

- capacity payments,
- operating reserves pricing,
- strategic reserve,
- capacity requirements,
- reliability contracts, and
- capacity subscriptions.

The generation market can be described by using the traditional economic variables of price and quantity. Capacity mechanisms can be grouped as to whether they leave all of these variables to the market, or whether one or both are determined by a central planning

agency (Jaffe and Felder, 1996). There is a relationship between the two variables and the goals of effectiveness and efficiency. Capacity mechanisms, this chapter argues, tend to be more effective when they make the demand for peaking capacity concrete. A capacity mechanism that requires a specific volume of available generation capacity from the market and provides financial compensation to the owners of that capacity reduces investment risk. Efficiency, on the other hand, is impacted by the degree to which investment decisions are made competitively and decentrally, rather than through central planning. Therefore capacity mechanisms that rely upon financial incentives can be expected to be more efficient, *ceteris paribus*.

Figure 6.1 ranks the capacity mechanisms that will be discussed along these variables: the horizontal axis shows the degree to which the demand for generation capacity is made explicit, while the vertical axis shows the degree to which the mechanism relies upon financial incentives. In the remainder of this chapter, the capacity mechanisms will be discussed in order of increasing explicitness regarding the need for peaking capacity, and from low to high reliance upon financial incentives.

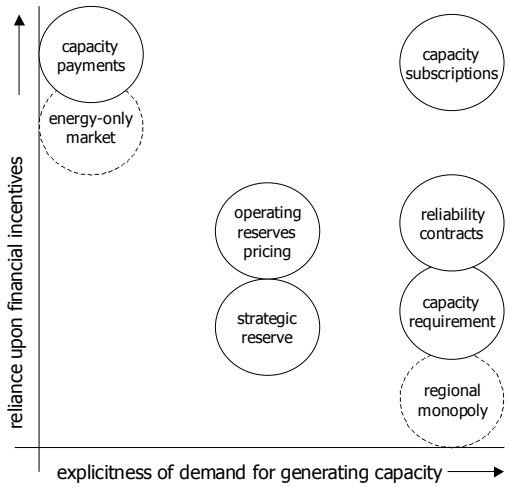


Figure 6.1: Categorization of capacity mechanisms

In energy-only markets, the demand for peaking capacity is the least explicit. It is up to the generating companies to estimate the likelihood that the current generating stock is insufficient to meet demand and to estimate the profitability of building new generating units to augment supply. The lack of transparency of the market and the lack of historical data about the statistical variability of supply and demand functions make this a difficult exercise, as was argued in Chapter 5.

Section 6.2 describes capacity payments, which are payments made to generating companies for installed or available generation capacity. The purpose is to shift the investment equilibrium to a higher volume, but this capacity mechanism does not provide a clearer indication of the demand for generation capacity.

Two solutions that are more capacity-oriented are a strategic reserve (described in Section 6.3) and operating reserves pricing (Section 6.4), in which the system operator stabilizes generator revenue through the way he contracts for operating reserves. Withdrawing this capacity from the market stimulates investment to meet regular demand. Total demand for peaking capacity is not made explicit but the investment risk is reduced by creating a separate demand for reserve capacity.

The last three options to be discussed make the demand for generation capacity fully explicit either by making it an administrative requirement or by revealing actual consumer demand for peaking capacity. In a system with capacity requirements, the regulator requires the market to provide a certain percentage of reserve capacity, as a result of which a capacity market is created (Section 6.5). The same principle underlies reliability contracts, as described in Section 6.6, but they have the advantage of providing better operational incentives. Arguably, the most market-oriented system is a system of capacity subscriptions, described in Section 6.7. In this system, a physical change in the market structure is used to create separate markets for electricity and for generation capacity.

When analyzing the different market designs, it should be kept in mind that the electricity price and the volume of generation capacity are not independent variables. One cannot influence the electricity price without having an impact upon investment incentives because investment is a function of the long-term expected average price (at least in a competitive market). On the other hand, measures that impact the volume of generation capacity will influence the price. While these variables are related, their relationship is not known precisely. Therefore, it is preferable to let at least one of them be determined by the market.

A general principle of capacity mechanisms is that they attempt to replace the investment incentive that is provided by price spikes with a more stable incentive, so that investment risk is reduced. This means that capacity mechanisms must create an alternate revenue source for generating companies. To the extent that investment risk is a function of inherent uncertainties regarding the availability of generation capacity and the stochastic nature of demand, investment risk is not actually reduced but shifted to consumers. (The degree by which this occurs depends upon the capacity mechanism.) This is justified because the consumers are the ones who can impact the investment risk, at least the part caused by demand fluctuations, and they are the ones who benefit from a higher volume of generation capacity.

The capacity mechanisms are designed such that generator income in each capacity mechanism equals the average income in a perfect energy-only market (except that the reduction of investment risk brought about by the capacity mechanisms may lead generating companies to require a lower risk premium). In practice, an increase in the end user price may be observed upon implementation if the current market price is below the long-run marginal cost of generation capacity. In the long run, this price increase would have been inevitable anyway to finance new capacity, and should be offset by lower price spikes and lower costs due to service interruptions.

When analyzing capacity mechanisms, it is important to realize that their design is based upon the adjustment of related variables. The principle of many capacity mechanisms is that they create a second revenue stream to generating companies, in addition to the revenues from electricity sales. Examples of this second income stream are capacity payments, the sales of reliability contracts, capacity credits and contracts to provide operating reserves. The second revenue stream can be created directly, for instance through capacity payments or reliability contracts, or indirectly, for example through capacity requirements. Where regulation affects these revenues streams, care should be taken that the sum equals the long-run marginal cost of the generating companies. Preferably, a sufficient number of variables is left to market forces to ensure that the generating companies can recover their costs at the optimal volume of generation capacity.

6.2 Capacity payments

Capacity payments provide generators with an economic incentive to maintain more capacity than they would otherwise. Generally, generating companies' expected revenues fall and costs rise with a higher volume of installed capacity. The idea behind capacity payments is that the additional revenues shift the investment equilibrium to a higher volume of generation capacity. A central planner determines a price, or a price function, which all generators receive in exchange for installed or available generation capacity. The lower net cost of generation capacity should lead generating companies to invest more. Capacity payments have been implemented in Columbia, Spain and Argentina (Vázquez et al., 2002).

Figure 6.2 shows the effect that capacity payments should have upon supply. The curves in the figure are hypothetical; the figure only serves to elucidate the concept of capacity payments. The demand curve varies continuously; future demand cannot be predicted exactly. As a result, future demand can only be characterized as a stochastic function. To keep the figure simple, only the average demand curve (D_a) and an example of a case of high demand (D_h) are shown. Figure 6.2 shows the supply curve that exists in an energy-only market as a solid line (S).

In the example of this figure, there will be a shortage of generation capacity when demand is as high as indicated by the high demand curve (D_h), which means that service interruptions occur. The purpose of capacity payments is to stimulate generators to invest more in generation capacity, which means the supply curve would be extended as indicated with the interrupted line (S'). With this longer supply curve, D_h can also be met so that the probability of a shortage of capacity has been reduced.

A question is how to set the level of the capacity payments. As they are intended to correct flaws in the market (underinvestment due to a lack of transparency or risk aversion, for instance), there is no theoretical justification for their level. It may be necessary to adjust the payment level in response to observed investment behavior; changes in the payment level entail a risk, however, of undermining their credibility as a long-term investment incentive.

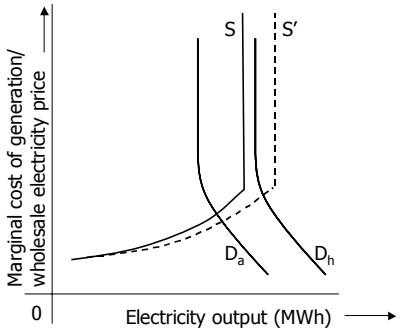


Figure 6.2: Capacity payments should lead to a larger volume of generation capacity and therefore extend the supply curve

A different kind of capacity payments were in the former England and Wales Pool (Wolak and Patrick, 1997). These payments varied depending on the reserve margin, hence their name 'dynamic capacity payments'. The payments were larger when the need for more capacity became more urgent. Their basic structure was:

$$capacity\ payment = LOLP \cdot (VOLL - max[SMP, bid\ price])$$

in which LOLP is the loss-of-load probability, VOLL is the average value of lost load and SMP is the system marginal price (which was the base for the pool price). As the England and Wales Pool was an integrated system, the reserve margin was known from hour to hour to the market (pool) operator so he could calculate the LOLP and, with that, the capacity payment. The payment was made both to active generators and to ones that were out of merit.

6.3 Strategic reserve

A strategic reserve consists of a set of generating units that are kept available for emergencies by an independent agent, typically the system operator. As these units are expected to operate only sporadically, the most economic way to establish such a strategic reserve is by purchasing old units from generating companies, hence the nickname 'mothball reserve'. However, the agent may be forced to construct new capacity if the market offers too little old capacity for sale. The reserve is paid for by revenues from a fee on electricity consumption, for instance a surcharge on system tariffs, and with the revenues from producing electricity for the market. In theory, the revenues from dispatching the reserve should equal the costs. It may be necessary, however, to finance the reserve through a different source. In order not to distort the generation market, the strategic reserves are only deployed when there is a shortage of electricity. The agent should be strictly neutral regarding the other players in the electricity market, and the deployment of its reserves should be limited to emergency conditions in order to minimize interference with the electricity market.

An important aspect of the design of a strategic reserve is the conditions under which the

strategic reserve is deployed. Two types of 'triggers' can be defined: a technical (capacity) and an economic (price) trigger. The first option means that the reserves will only be deployed if the generation capacity that is offered on the market is insufficient, or nearly insufficient, to meet demand. The second option is to deploy the reserves when the electricity price has reached a certain level. In both cases the question needs to be answered as to the selling price of electricity from the strategic reserve.

Using a 'technical' trigger would mean that the strategic reserve would be deployed any time the margin between available generation capacity and demand would fall below a certain level. If electricity from the reserve would be offered to the market at the marginal cost of generation, this would severely distort the generation market, as it would eliminate much of the scarcity revenues that generators might otherwise expect. This would have a strongly depressing effect upon investment in generation capacity. In order not to distort the generation market, the choice may be made to offer the reserve capacity at the price of the highest supply bid on the market. This would ensure that no commercial generators would be displaced but appears highly susceptible to manipulative bidding.

The other option is to establish a fixed price P_{sr} at which the strategic reserve is deployed. P_{sr} becomes the de facto maximum price for the market; the lower this price is, the smaller the incentive will be for private parties to invest. The volume K_{sr} of capacity that an agent needs to maintain as a reserve therefore depends upon P_{sr} (and, of course, upon the level of reliability that is desired). Therefore the volume of the operating reserves and the dispatch price P_{sr} should be adjusted to each other. A fundamental shortcoming is that the administrator of this system needs to determine both the price P_{sr} and quantity K_{sr} of the reserves, while their exact relation may be difficult to know, as it depends upon the shape (especially the top end) of the load-duration curve.[28]

A strategic reserve creates an elastic section at the end of the supply curve. In Figure 6.3 the strategic reserve is indicated by the horizontal section in the supply curve *S*. The electricity from the reserve is sold at a price P_{sr}. The figure shows the average demand curve D_a and a high demand curve D_h.

If the strategic reserve is to function only as a back-up system, the theory of an energy-only market should be applied and P_{sr} should equal the average value of lost load. In that case, the reserve should not be expected to be deployed except during rare moments. A different approach is possible, however. It may be chosen to deploy the strategic reserve at a lower price P_{sr}. The reduced incentive to invest in generation capacity means that the competitive generating companies would provide a smaller volume of generation capacity. This could be compensated by maintaining a larger strategic reserve K_{sr} which would operate more frequently. The theoretically optimal price P_{sr} at which to dispatch the strategic reserve depends upon the volume of generation capacity that is to be provided by the market. P_{sr} can be determined if the price-duration curve, and its stochastic distribution, is known. The more generation capacity the market is expected to

[28] This problem also exists with operating reserves pricing, but not with capacity requirements, reliability contracts and capacity subscriptions. The latter three solutions, we will see, leave the price or both variables to the market.

provide (the smaller the strategic reserve), the higher P_{sr} should be.[29] Example 6.1 provides a sample calculation which illustrates this relationship.

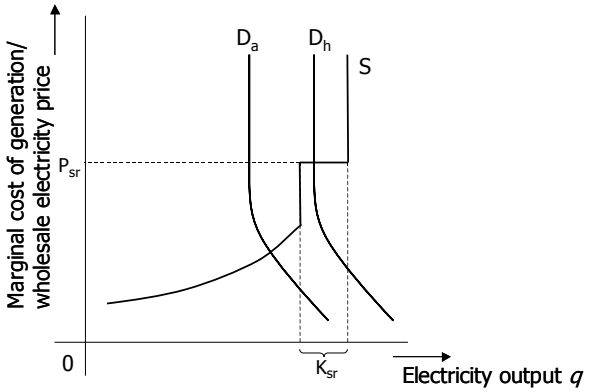


Figure 6.3: A strategic reserve introduces a perfectly price-elastic section into the supply curve

The new European Electricity Directive allows member states to enter a 'tendering procedure' for new generation capacity in the case that existing generation resources appear insufficient (Art. 7, Directive 2003/54/EC). Depending upon the details of the measures taken, the tendering procedure may resemble the creation of a strategic reserve or capacity payments. If the tendering procedure simply is a way to stimulate the construction of new capacity, it may be regarded as a capacity payment. However, this would have a distorting impact upon competition and investment incentives. If the payments to generators are accompanied by conditions regarding the availability of the generating units and the price at which they sell electricity, the procedure may be more similar to a strategic reserve. The vagueness of the Directive prevents a more specific analysis. However, as it will probably resemble one of the systems described in this chapter, the tendering procedure will not be analyzed in further detail here.

Example 6.1: Strategic reserve

If a strategic reserve of generation capacity is maintained with the purpose of deploying it during periods of scarcity, its price must be calculated carefully to ensure that the reserve does not depress the incentives for sufficient investment in generation capacity. While the reserve only provides power during peak conditions, the price at which the reserve is dispatched impacts the average revenues that commercial generating companies receive and therefore the level of investment. First, the optimal volume of generation capacity will be determined. Given a choice for a certain volume of reserve

(Example continued on the next page.)

[29] This reasoning is taken from Stoft (2002), who applies it to operating reserves pricing. See also the next section.

(Example continued from the previous page.)

capacity, the optimal dispatch price of the reserve in order not to distort the investment signal can then be determined.

In an equilibrium, the market should provide K_M of generation capacity, so together with the capacity of strategic reserve K_{sr} the optimal volume of generation capacity K^* is obtained:

$$K_M + K_{sr} = K^* \qquad (6.1)$$

The optimal volume of available generation capacity K^* can be estimated from the average value of lost load and the load-duration curve. The optimal volume of generation capacity is reached when the long-run marginal cost of new capacity equals the average value of lost load (VOLL). At this point, the duration (number of hours per year) that load is shed $f_d(q)$ is optimal. (See also Section 5.4.) The long-run marginal cost of generation capacity can be approximated by the annual fixed cost C_F of a peak plant, which will be assumed to be 40,000 €/MW (Newbery et al., 2003), divided by the number of hours that it operates. This number, in turn, is approximately equal to the average number of hours per year with insufficient generation capacity. The average value of lost load in the Netherlands was recently estimated to be about 8,600 €/MWh (Bijvoet et al., 2003). Then, using equation (5.9)

$$f_d(q)^* = C_F/V_{ll} = 40{,}000/8{,}600 = 4.7 \text{ hours/year.} \qquad (6.2)$$

Under the assumptions made, the optimal average duration of load shedding $f_d(q)^* = 4.7$ hours per year.

If the load-duration curve is known, the volume of available generation capacity that is required to meet this level of reliability can be calculated. This is not so easy, however, because the top part of the load-duration curve often is not well known. Moreover, for a proper calculation, it is not sufficient to have a single historical (or average) load-duration curve because the load-duration curve varies stochastically. Finally, the stochastic variations in volume of available generation capacity also impact the probability that the marginal generator is called upon.

For the sake of this example, let us assume a simple load-duration curve similar to the one that Stoft (2002) uses as an example. The stochastic nature of the availability of generation capacity will be ignored. Generator outages will be incorporated in the

augmented load L_g, a term introduced by Stoft (2002) that consists of load q plus generator outages. To mimic the stochastic nature of the load-duration curve, assume there are only two realizations, one for normal years (3 out of 4 years on average) and one for the remaining 'hot' years:

$$f_d(L_g) = 4 \cdot (19 - L_g)^2 \text{ during hot years (1 in 4 years)} \qquad (6.3a)$$

$$f_d(L_g) = 4 \cdot (16 - L_g)^2 \text{ during normal years (3 in 4 years)} \qquad (6.3b)$$

(Example continued on the next page.)

(Example continued from the previous page.)

Assume that load shedding only needs to take place during hot years. During those years, the optimal duration of load shedding is 4·4.7 h/y = 18.8 h/y. Substituting this in (6.3a) renders a corresponding augmented load L_g of 16.8 GW. Total installed capacity K^* must equal L_g:

$$K^*=16.8\ GW^{30}$$

If a strategic reserve of 1 GW is maintained, the market must provide 15.8 GW. To determine the correct dispatch price for the strategic reserve P_{sr}, first it will be calculated how often this price is reached. By substituting L_g=15.8 GW in (6.3a) and (6.3b), we find that the reserve is called upon for 41.0 h/y during hot years and for 0.2 h/y during regular years. On average, this is 10.4 hours per year. This means that the marginal generator that is not in the reserve also runs about 10.4 h/y on average. Assuming fixed costs of 40.000 €/MW per year, the generator needs an average price of 3,846 €/MWh to recover its fixed costs. This should therefore also be the dispatch price P_{sr} of the reserve.

If the size of the reserve is increased to 2 GW and the market only needs to provide 14.8 GW, we find that the reserve is called upon at an average of 22 hours per year. The associated P_{sr} = 1,818 €/MWh. The larger the reserve, the lower the dispatch price can be without distorting the investment incentive.

This example shows that in order to deploy a strategic reserve without distorting the investment incentive for generating companies, detailed knowledge is required:

- The load-duration curve and its stochastic distribution must be known;
- The average value of lost load must also be known in order to estimate the optimal volume of generation capacity.
- The stochastic distribution of the available generation capacity also needs to be known. (This aspect was not part of the example).

With this information, a planner first calculates the optimal volume of generation capacity, then decides the reserve volume and calculates the optimal dispatch price. The reverse is also possible: given a certain reserve dispatch price P_{sr}, the optimal reserve volume can be calculated.

6.4 Operating reserves pricing

Every electricity system needs operating reserves to maintain the physical stability of the system. The system operator uses these reserves to manage the difference between expected and actual demand. Stoft (2002) proposes a systematic way of paying for these reserves that stabilizes generator revenues and that should work as follows. The system operator sets as a goal to permanently contract a certain volume of generation capacity,

[30] If the necessary data to make this calculation cannot be obtained, the desired volume of generation capacity may be determined on the basis of forecasts of the future development of peak load and a certain reserve margin. In the absence of sufficient data, this is a somewhat arbitrary guess, often based upon a sense of what historically has been perceived as reasonable. The possible sub-optimal price-quantity combinations which may result are a reason why the capacity mechanism may not be revenue-neutral.

for instance 10% of peak demand. He contracts this capacity, for instance in a daily auction, and pays for its availability, even if it does not operate.[31] The system operator is only willing to pay a certain maximum price for reserve capacity. This means that during a period of scarce capacity, the system operator may contract less capacity than his target. Thus, the capacity that is contracted as operating reserves during off-peak periods becomes available to the market during peak demand. It becomes available at an electricity price that corresponds to the system operator's willingness to pay through arbitrage of the two markets (the spot market price should be somewhat higher to account for the variable costs of electricity production). As a result, the system operator's maximum purchasing price P_{or} becomes a *de facto* system price cap. Every time that demand would push the price for generation capacity beyond the system operator's price limit, the system operator would contract less capacity, reducing demand and stabilizing the price. However, if all generation capacity is exhausted the price may still rise beyond P_{or} up to the average value of lost load.

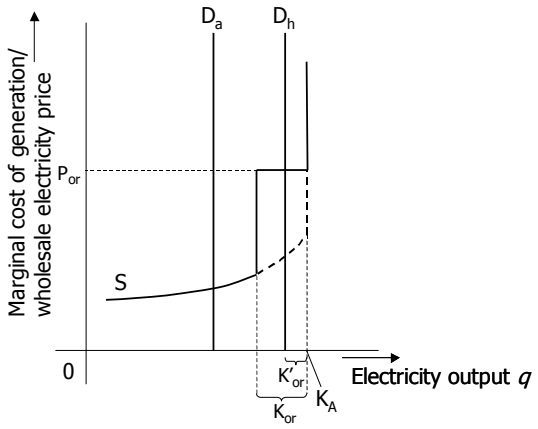


Figure 6.4: Operating reserves pricing raises the price of the last section of the supply curve

The effect of operating reserves pricing is that the system operator raises the price of the last section of the supply curve. As a result, the electricity price starts to rise well before actual scarcity occurs, so an earlier investment signal develops. Consequently, more but smaller price spikes develop, which stabilizes generator income and contributes to the predictability of market prices (see Figure 6.4). The Y-axis shows the electricity price and the marginal cost of generation; the X-axis shows the volume of electricity produced (q). The total volume of available generation capacity is equal to K_A. The system operator purchases a volume of reserve capacity equal to K_{or} and is willing to pay at most P_{or} per unit of reserve capacity.

[31] This is different from some operating reserves markets, such as the Dutch market, where generators have an obligation to offer unused capacity but are only remunerated if they are dispatched. (Recently, the system operator has entered into some long-term contracts for operating reserves, so the Dutch system actually is a hybrid.)

Under normal conditions, the electricity price is determined by the intersection of the demand curve (indicated by the demand curve D_a, for average demand) and the supply curve (indicated with S). As there is an excess supply of generation capacity, the price of reserve capacity is close to the marginal cost of capacity (the costs of maintenance and of keeping the generating units stand-by). When electricity demand exceeds K_A-K_{or} (indicated by the demand curve D_h) there is not sufficient generation capacity to satisfy both the demand for electricity and the system operator's demand for operating reserves. Because the system operator is willing to pay no more than P_{or}, generating units that normally are sold as reserve capacity to the system operator may now find it more attractive to sell in the spot market. Consequently, the system operator purchases a smaller volume of reserves, equal to K'_{or} in Figure 6.4. The price stays constant until the generation capacity that normally is sold as operating reserves is all used for electricity production.

Figure 6.4 actually shows a simplified version of Stoft's proposal. Rather than capping the system operator's willingness to pay at a certain price, Stoft (2002) argues for a sloping demand curve. Thus, the system operator's willingness to pay for operating reserves would increase as the availability of reserves decreases. This would better reflect the value of operating reserves (and therefore be more efficient). More importantly, it would mitigate market power along a broader range of electricity prices because a reduction of generation capacity would no longer lead to steep price increases.

The system operator needs to carefully choose the right combination of the volume of operating reserves and maximum price that he is willing to pay, as these two variables together determine the revenues for the generators and hence the incentive to invest in generation capacity. As in the case of a strategic reserve, the designer of a system of operating reserves pricing needs to set both a price and a quantity variable. For this purpose he needs to know the load-duration curve, from which he can determine the expected operational time of the marginal generating unit that is not within the operating reserve. This unit should just be able to recover its costs at the operating reserve price P_{or}. If P_{or} is too low, this unit will eventually disappear from the market, reducing generation adequacy; if P_{or} is too high, the price spikes will be higher than necessary.

The dynamic effects of operating reserves pricing are more frequent but lower price spikes, as they are limited by the system operator's willingness to pay. During periods of abundant generation capacity, the value of reserve capacity would still be close to zero. The system operator's demand for capacity, however, leads to periods in which the available capacity is insufficient to meet this demand. Then the market price rises to the system operator's willingness to pay. The larger the reserve, the lower the system operator's willingness to pay needs to be and the more frequently the price will equal it, as is demonstrated in Example 6.2. Thus, a larger operating reserve will lead to a higher predictability of prices and therefore better stabilize investment. However, a higher reserve should be accompanied by a lower price P_{or} in order not to provide an incentive to over-invest. Too low a P_{or}, however, may suppress the incentive for price-elastic demand behavior. A variation of this system currently is being used by Statnett in Norway. There, the system operator buys options for reserve capacity from generators as well as interruptible contracts from consumers

Example 6.2: Operating Reserves Pricing

This example describes a simplified process for selecting the parameters for a system of operating reserves pricing. The same general setting will be used as in Example 6.1: an optimal volume of generation capacity K_T = 16.8 GW and an optimal duration of load shedding $f_d(q)^*$ = 4.7 h/y. The example will show that the calculation of the basic parameters for operating reserves pricing is similar to the calculation for a strategic reserve.

Our goal is to calculate P_{or}, the optimal willingness to pay for operating reserves, which becomes the *de facto* market price cap. First it will be assumed that an operating reserve volume of 1 GW has been chosen. The market price is equal to P_{or} when

$$L_g > K^* - K_{or} = 16.8 - 1 = 15.8\ GW \qquad (6.4)$$

So the electricity price is equal to P_{or} when $L_g > 15.8\ GW$. Using (6.3a) and (6.3b) from Example 6.1, it can be determined that this price will be reached 10.4 hours per year on average. (The weighted average of (6.3a) and (6.3b) must be used.) The marginal plant provided by the market will operate only these 10.4 hours per year and must recover its capital cost during these hours. Again assuming fixed costs of 40,000 €/MW per year (Newbery et al., 2003), the optimum for P_{or} can be calculated:

$$P_{or} = C_F / D_{or} = 40{,}000 / 10.4 = 3{,}846\ €/MWh \qquad (6.5)$$

Thus, for an operating reserve of 1 GW, the system operator must be willing to pay up to 3,846 €/MWh in order to provide an efficient investment incentive to the market.

If the operating reserve is expanded to 2 GW, P_{or} is reached 25.4 hours per year on average. Then P_{or} needs to be only 1,818 €/MWh to allow the marginal generator to recover its cost. These calculations are the same as for a strategic reserve.

It is important to realize that these calculations are highly stylized. In reality, prices will not only exceed marginal costs during electricity shortages. Rather, they will start to rise above marginal costs as a shortage approaches (Newbery et al., 2003). See for example Figure 4.6. Consequently, there are more hours during which generators recover their fixed costs, so the height of the price spikes may be lower. The system operator faces the challenge of estimating the magnitude of this effect, so he can adjust P_{or} accordingly.

This sample calculation teaches the following about the design of a system of operating reserves pricing:

- a planner calculates the desired total volume of installed generation capacity;
- this depends strongly upon the assumed fixed costs of the marginal generator, the average value of lost load and the load-duration curve;
- the load-duration curve must also be known in order to determine the optimal maximum price that the system operator should pay for operating reserves.

These requirements are the same as for a strategic reserve.

6.5 Capacity requirements

The system called capacity requirements is also known as a capacity market; however, the name 'capacity requirements' better describes its most characteristic feature.[32] Capacity requirements are part of the PJM Interconnection electricity system, one of the largest electricity systems in the world, where they are called ICAP for Installed CAPacity (PJM Interconnection LLC, 2003). Similar systems are in use in New York and New England (New York ISO, 2002).

In a system with capacity requirements, a central planning agency determines the desired generation capacity margin. Based upon the expected total coincident peak demand of the loads served by each load-serving entity (retail company or large consumer), the system operator calculates how much generation capacity each load-serving entity must purchase (PJM Interconnection LLC, 2003).[33] Reserve capacity may take the form of available generation capacity or interruptible contracts. Generating companies may sell capacity credits up to the volume of generation capacity that they have reliably available, which is determined by the regulator. Capacity credits can be traded, so there is a secondary capacity market. Load-serving entities include the cost of purchasing capacity credits in the price they charge to final consumers for electricity. In theory, if the capacity margin is chosen optimally, the average price paid by consumers should be the same as in a perfect energy-only market. The requirement to contract generation capacity in excess of the projected peak causes the capacity market to become constrained before the energy market does. As a result, the incentive to invest in new generation capacity develops before the electricity market becomes constrained.

If an energy-only market would install a volume K_O that would be deemed insufficient, the regulator could apply a capacity requirement of K_A. See Figure 6.5. This would induce investment in generation capacity, so the supply curve would be extended from S_E to S_E', as a result of which the probability of shortages would be reduced. In the figure, this is indicated by the intersection of the high demand curve D_h with the supply curve S_E'. The lower generator revenues in the wholesale market, which result from the greater availability of generation capacity, are compensated by the revenues from selling capacity credits (depicted on the right side of Figure 6.5). As the price of the capacity credits is determined in a competitive market, generating companies have the opportunity to fully recover their costs.

If available generation capacity is less than the total capacity requirement (indicated by K_A'), the supply of capacity credits (S_{CC}') will fall short of the load-serving entities' demand. This will cause some of the load-serving entities to default on their obligations, as a consequence they will need to pay the penalty P_{pen}. Consequently, the penalty

[32] The description of this method is largely based upon Doorman (2000), PJM Interconnection, L.L.C. (2001) and Hobbs et al. (2001c).

[33] In principle, the capacity requirements could also be placed on other parties, such as generating companies or consumers. Using the load-serving entities appears most practical, however. A disadvantage of placing the requirement upon generating companies is that the trade of the capacity credits may be affected by strategic behavior, while placing the requirement upon consumers would create large transaction costs.

becomes the upper limit to their willingness to pay for capacity credits, so it functions as a maximum price in the capacity market. P_{pen} must be chosen carefully, as it determines the incentive to invest in generation capacity. At a minimum, the penalty must exceed the cost of new plant (Shuttleworth et al., 2002).

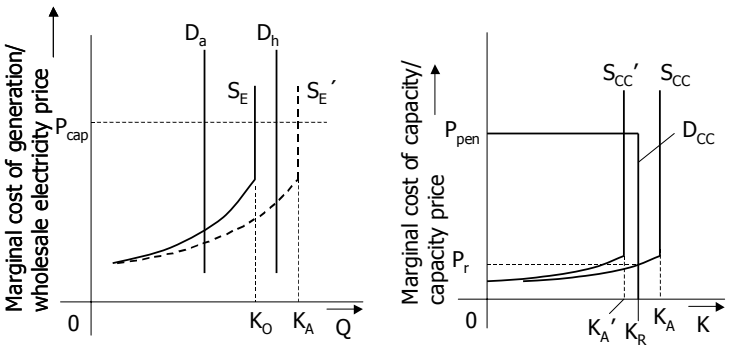


Figure 6.5: Capacity requirements increase the volume of installed capacity (left) and create a separate capacity market (right)

The capacity requirement may cause the capacity market to be volatile because the demand for the capacity credits (D_{CC} in Figure 6.5) is perfectly price-inelastic. If the available volume of generation capacity K_A is larger than the capacity requirement K_R, the price of the capacity credits is determined by the marginal cost of maintaining reserve capacity, indicated by the supply curve for capacity credits S_{CC} in the right-hand part of Figure 6.5. The resulting capacity credit price is low (P_r in the figure). When the supply of generation capacity is insufficient to meet the demand for capacity credits (when $K_A<K_R$), the capacity price is determined by P_{pen}. As a result, small changes in the availability of generation capacity may cause the capacity price to oscillate between the marginal cost of providing reserve capacity and the penalty price.

When the PJM market is short of electricity, the system operator 'recalls' generation capacity. All generators that have sold capacity credits are required to offer their capacity into the PJM pool, even if they have export contracts. Thus the capacity requirement is a type of call option, with the strike price equaling the pool price cap P_{cap} ($1,000/MWh). Capacity requirements can (and arguably should) be combined with an energy price cap because price spike revenues are not needed to finance peak capacity. Stoft (2002) argues that without a price cap, the system would in theory even lead to over-investment. More importantly, a price cap be instrumental in mitigating market power.

Example 6.3: Capacity requirements

The same conditions will be continued to be used as in the previous two examples. A planner uses the average value of lost load, the load-duration curve and an estimate of the long-run marginal costs of generation to estimate the optimal volume of generation capacity K^*. This was calculated to be 16.8 GW in Example 6.1. The total capacity requirement $K_{R,T}$ is set equal to the optimal volume of installed capacity. As was mentioned in Example 6.1, in the absence of sufficient data the above process may give way to a more prosaic method. In PJM, forecasts of the expected development of peak demand are combined with a reserve margin to arrive at the desired total volume of generation capacity.

Assume a planning period of 5 years, and an expected average growth rate of 2% per year. The expected peak demand in 5 years D_5 then is:

$$D_5 = 1.02^5 \cdot 16.8 = 18.5\ GW$$

If a reserve margin of 17% is applied, the total desired volume of available generation capacity K_A would be:

$$K_A = 1.17 \cdot 18.5 = 21.7\ GW$$

Available capacity is less than installed capacity. If on average 8% of generation capacity is unavailable, the installed capacity requirement is adjusted accordingly:

$$K_I = \frac{K_A}{1 - 0.08} = 23.5\ GW$$

In PJM, the capacity requirement is adjusted for the participation of interruptible load.

The regulator distributes the installed capacity requirement among the load-serving entities in proportion to their market shares. The load-serving entities are required to purchase capacity credits from generating companies equal to their individual capacity requirements. The price of the capacity credits is determined competitively by the generating companies (the suppliers of the credits) and the load-serving entities (who buy them).

Example 6.1 showed that if the marginal generator has fixed costs C_f of 40,000 €/MW per year would run 4.7 h/y if the volume of generation capacity were optimal. It would need to receive a price equal to the average value of lost load, 8,600 €/MWh, in order to simply recover its costs. Assume the price cap P_{cap}=1000 €/MWh (in PJM it is 1000 $/MWh). The generator will need to recover the income that it loses due to the price cap in the capacity credits market. Its lost income is equal to:

$4.7 \cdot (8600 - 1000) = 35{,}720\ €/MW$ per year.

The average price of capacity credits would need to equal 35,720 €/MW per year, or 4.08 €/MW per hour, for the marginal generator to break even.

Generating companies higher in the merit order would typically have higher fixed costs, so unlike the marginal unit they would not recover all their fixed costs in the capacity credit market. However, their variable costs are lower than the price-setting generator for part of the time, so that they make an operating profit.

6.6 Reliability contracts

Reliability contracts are designed as an improvement upon capacity requirements.[34] They provide generators with better incentives for making their resources available during periods of scarce supply. An independent agent purchases call options from generators on behalf of consumers. The call options give the agent the right to the difference between the electricity spot price P_m and the option strike price. This price difference is then returned to consumers, so that the net amount they spend during price spikes is limited by the option strike price P_s.

Let us assume that the system operator is also the agent who operates the capacity mechanism: he purchases the options, calls them and redistributes any proceeds to the consumers. The volume of the contracts and the strike price are determined by a central planner. The volume of reliability contracts is equal to the forecasted coincident peak load plus a reserve margin, similar to in a system with capacity requirements. The strike price should be above the highest marginal cost of operation of all the generators, to make sure it will not discourage any generator from producing. The price of the reliability contracts (the option premium) is determined in auctions.

The system operator calls the options any time that the market price exceeds the option strike price. Then generating companies who have sold options pay the system operator P_m-P_s times the volume (in MW) for which they have sold options. An operational generator will receive P_m from selling electricity on the market, so his net income will be equal to P_m minus his payment (P_m-P_s), which is equal to P_s. The generating company's option payments are fully hedged by market prices, as the generation capacity that backs the option contracts is operational.

A generator who has sold option contracts but happens to be unavailable when the options are called, still is required to pay (P_m-P_s) – but does not have any revenues to compensate these payments. In this case the payments cause a net loss. Therefore generating companies have a strong incentive to make their capacity available when the options are called, which is when electricity is scarce.[35] This is one of the main advantages of this system. A second advantage is that the generating companies have an incentive to sell a volume of call options equal to their expected output: selling less would lower their revenues, while selling more would expose them to a high price risk during shortages.

For consumers, the effect is that the system operator has 'purchased' a price cap equal to P_s. As this limits the average revenues of generating companies, the latter will demand a price for selling the option contracts that corresponds to the expected loss of price spike revenues, which is the sum of (P_m-P_s) over all hours that P_m>P_s. As the option price is determined by the generation market in a competitive auction, it should reflect

[34] The description is based upon Vázquez et al. (2002), who proposed this system. Oren (2000) outlined a similar proposal, which will be discussed in Section 8.2.

[35] The proposal by Vázquez et al. (2002) adds a fixed penalty to the payments by the generators to further discourage them from not being available. An attractive consequence is that reliable generators are able to bid lower in the auction than unreliable generators.

generators' expected price spike revenues. If the system functions well, there should be no net cost to consumers, as the cost of extra generation capacity should be offset by the reduction of price spikes and the benefit of increased reliability of service.

Figure 6.6 shows that the use of reliability contracts results in an extension of the supply curve S_E to S_E', the same as capacity requirements do. In case of shortages, the net electricity price paid by consumers is capped by the option strike price P_S.[36] Figure 6.6 is similar to Figure 6.5 (recall that capacity requirements may be combined with an energy price cap). From a theoretical perspective, the systems are quite similar: a regulatory body determines the total desired volume of generation capacity, strong financial incentives are used to obtain this volume, and the market determines the price of capacity. In exchange for payments for available capacity, energy prices are limited. In theory, the systems should have the same effects upon investment, system efficiency and welfare. Implementation details and operational incentives are quite different, however, as will be seen in the evaluation in Section 7.7. One significant difference is that the reliability contracts provide a strong incentive to maximize generator output when capacity is scarce.

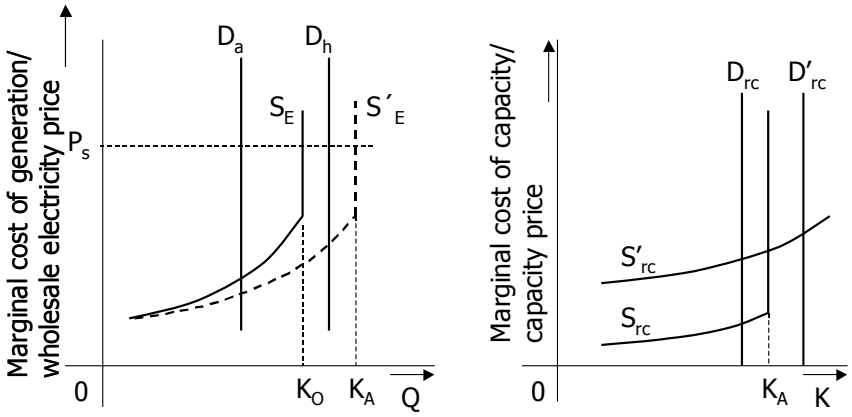


Figure 6.6: Reliability contracts extend the supply curve (left); the contract price is determined in an auction (right)

[36] In principle it is possible that there is generation capacity available that is not covered by the reliability contracts, as the generating companies may be conservative with their estimates of the availability of generation capacity and adjust their sales of reliability contracts accordingly. However, the probability that prices rise far above the marginal cost of generation when excess generation capacity is available is slim, as this could only be caused by an increase in demand in excess of the capacity margin. In this situation, a high price spike could occur, but the consumers would be hedged against nearly the entire volume of their electricity consumption, so it would not affect their average price by much.

Example 6.4: Reliability contracts

The total demand for reliability contracts is equal to the forecast optimal volume of available generation capacity. It can be determined in the same way that the capacity requirement was determined in Example 6.3 but without correcting for the outage rate because the reliability contracts represent available capacity rather than installed capacity. Each generating company decides the volume of reliability contracts it sells based upon its installed capacity and its outage rate. The generating companies have an incentive to minimize their outages, as this allows them to sell a larger volume of reliability contracts.

The system operator determines the strike price P_s of the reliability contracts. The choice of P_s is a trade-off between the stabilizing effect of lower prices (they reduce price volatility) and the incentive for load reduction created by higher market prices. Let us assume the system operator has chosen P_s=1000 €/MWh, the same as the price cap in Example 6.3. This means that during episodes of scarcity, the generating companies lose price spike income. If they could have earned a price equal to the average value of lost load (which had been assumed to be 8,600€/MWh), they lose 7,600 €/MWh during shortages.

Now let us assume that the system works as intended and the duration of load shedding is optimal. Example 6.1 calculated the optimal duration of load shedding to be 4.7 hours per year. In an energy-only market, the electricity price would have risen to the average value of lost load during these hours. If it is assumed that the electricity price was close to the marginal cost of operation the rest of the time, the income that the generating companies lose from selling reliability contracts with a strike price P_s can be calculated. This revenue loss is equal to

$$4.7 \cdot 7{,}600 = 35{,}720 \ €/MW \text{ per year.}$$

The generating companies would recover these revenues through the sales of the reliability contracts; therefore it is to be expected that the price of these option contracts would also be around 35,720 €/MW per year. Deviations may point to excess capacity or a projected capacity shortage; the long-run average may also be somewhat lower if the system indeed leads to a reduction of investment risk, as it is intended to do. The expected costs of the reliability contracts is the same as the expected costs of capacity credits in a system with capacity requirements, which is not surprising as the principle underlying both capacity mechanisms is the same.

The above figure is on the order of magnitude of the long-run marginal costs of a generator. This is not surprising because in a perfectly competitive system with reliability contracts, generating companies can recover only a small part of their fixed costs during periods when supply exceeds demand. In an energy-only market, they would have recovered their fixed costs during the price spikes.

As with capacity requirements, there should be a penalty for failing to deliver. Here, however, the penalty is determined by the market: the greater the shortage, the higher the penalty. A generator that has sold reliability contracts but is unable to produce electricity when the contracts are called, is liable to pay the difference between the market price and the option strike price, in this case up to 7,600 €/MWh if the market price rises to the average value of lost load.

(Example continued on the next page.)

(Example continued from the previous page.)

For an individual generating company with, for instance, 1000 MW of installed capacity, the system works as follows. The company estimates the availability of its generating units. Suppose it concludes that most of the time, at least 90% of the units is available. It would then sell an equivalent of 900 MW of option contracts to the system operator. If the price indeed is 35,720 €/MW per year, this yields an annual revenue stream of 32,148,000 €.

Most of the time, the company produces and sells electricity like in an energy-only market. When the electricity price exceeds the strike price P_s, however, its options are called. This means that the company pays the difference between the market price and the option price to the system operator. Suppose the strike price is 1000 €/MWh and the electricity market price is 2000 €/MWh. The generating company then pays the system operator the following amount:

$$payment = (2000 - 1000) \cdot 900 = 900{,}000 \ €/h$$

If the generating company produces at least 900 MW, the payments are more than offset by the revenues from selling electricity on the market. The net effect of the payments is to cap the generating company's revenues at 1000 €/MWh for the 900 MW for which it has sold option contracts. A higher market price leads to higher option payments but also to higher revenues to the same degree, so the net revenues remain 1000 €/MWh for the 900 MW for which call options were sold. Only the remaining 100 MW creates net revenues equal to the full market price. Thus, if all units are available the generating company's revenues are:

$900 \cdot 1000 = 900{,}000 \ €/h$ from the units that underlie the options, plus

$100 \cdot 2000 = 200{,}000 \ €/h$ from the remaining units.

So the net revenues are 1,100,000 €/h. The same result can also be found differently. Gross revenues are equal to output multiplied by the market price 1000 MW·2000 €/MWh=2,000,000 €/h; net revenues are the difference between gross revenues and the option payments of 900,000 €/h (calculated above).

The payments need to be made regardless of the actual output of the company's generators. Therefore the company has a strong incentive to maximize its output, as each additional unit yields a revenue of 2000 €/MWh – and vice versa, each unit less output means a reduction of revenues by that amount. (The marginal revenues are the same whether the generating company produces more or less than 900 MW.) Thus, generating companies have an incentive to maximize output while the sales of the reliability contracts stabilizes their net revenues.

The generating company also has an incentive to limit the sale of options to the volume of generation capacity that it expects to have available because options that are not covered by electricity sales constitute a significant financial risk. As a result, the total volume of option contracts offered to the system operator should constitute a good estimate of available generation capacity.

The payments that the generating companies make to the system operator are transferred to the consumers. Consequently they have the benefit of a de facto price cap equal to the strike price P_s. The system dampens the price volatility, which benefits both consumers and producers.

Another important difference with capacity requirements is the way the capacity market works. In the case of reliability contracts, there is a single buyer who purchases the contracts in recurring auctions. In times of excess generation capacity (available generation capacity K_A exceeds the system operator's demand for reliability contracts D_{rc}), the price of the reliability contracts is determined by the short-term marginal cost of maintaining reserve capacity (S_{rc} in the right-hand side of Figure 6.6), as in a market with a capacity requirement. Scarcity prices need not develop in the auctions, however, if they are held long enough in advance such that participants may place bids that are to be covered with generation capacity that has not yet been constructed. In that case, the auction price is determined by the long-run marginal cost of generation (indicated by S_{rc}' in the figure).

6.7 Capacity subscriptions

A market with capacity subscriptions is the most market-oriented of the solutions discussed here, as both the quantity of reserve capacity and the price are determined by the market.[37] Again, a separate market for capacity is created. Consumers are required to buy from generators the right to the amount of capacity that they wish to have reliably available to them during peak demand periods. (During off-peak times, consumption is not limited.) By forcing consumers to pay for the generation capacity that is made available on their behalf, generators receive a signal regarding the demand for generation capacity and therefore are induced to provide the amount of capacity that consumers consider optimal. This way, consumer preferences for reliability are correctly reflected in the volume of available capacity. This is the only capacity mechanism discussed so far in which consumers can directly influence the volume of installed generation capacity, like they do in an ideal energy-only market. (See Section 5.2.1.)

A system of capacity subscriptions works as follows. Consumers need to install a sort of electronic fuse, which normally is not active. When the demand for electricity begins to approach the available generation capacity, the system operator activates the fuses. Then each consumer's electricity use is limited to the capacity of his fuse. Consumers can choose the size of their fuses; the price of a fuse depends on the cost of the generation capacity that is needed to guarantee a peak supply of electricity equal to the size of that fuse. The payments made for the fuses represent the costs of keeping an equivalent of generation capacity available, while the price of electricity represents the variable cost of electricity production. Consumers who want to be able to consume much electricity during peak hours need to pay more than consumers who do not mind shifting their consumption to another time. Thus, incentives are introduced for consumers to manage their own loads and rationing occurs in an economically efficient manner.

In Figure 6.7, the total volume of capacity subscriptions is indicated by K_a, which is equal to the generators' estimate of maximum available generation capacity. Because consumers purchase as much capacity from generators as they ever expect to need at one time, generators are forced by contract to install sufficient capacity to meet total

[37] The description of this solution is based upon Doorman (2000).

coincident peak demand. Therefore the supply curve is extended, as is indicated by the curve S_e'.

When demand does not exceed available capacity, the fuses are not activated and the market is not affected, as is indicated by the curve D_a in Figure 6.7. However, when demand is so high that the supply and demand functions would not intersect, the system operator activates the fuses. Thus he limits demand to the physically available volume of generation capacity K_a.

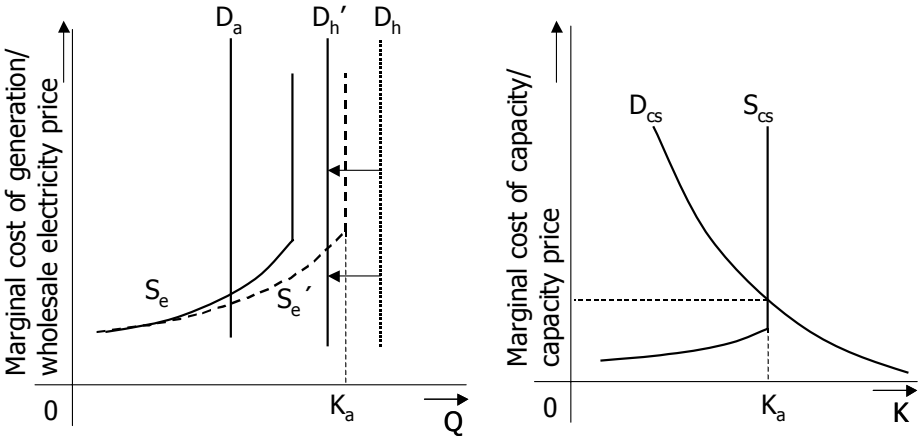


Figure 6.7: A system of capacity subscriptions extends the supply curve (left); the capacity subscriptions are traded in a separate market (right)

The capacity subscriptions can be traded in a separate market, depicted on the right side of Figure 6.7. While in the short term the supply of capacity subscriptions S_{cs} is limited to the available volume of generation capacity K_a, price volatility is limited by the fact that the demand for capacity subscriptions D_{cs} will be much more price-elastic than the demand for electricity itself. As the price of the capacity subscriptions rises, some consumers will prefer a smaller fuse over the higher cost. When it rises above the long-run marginal cost of generation, the capacity subscriptions market will attract new investment.

A complication is the stochastic nature of the availability of generation capacity, which means that generating companies need to sell less capacity than they have installed. However, even if a generating company maintains an ample margin between his installed capacity and the volume of capacity subscriptions that he sells, there is a possibility that he will not be able to meet his obligations. The first recourse is the balancing market but there is a remaining probability that the entire system is short of available capacity compared to the volume of capacity subscriptions that have been sold. This means that service interruptions may still occur. To keep these to a minimum, generating companies who do not meet their obligations should pay a penalty equal to the average value of lost load. The penalty can be paid to the consumers whose service was interrupted as a compensation.

Example 6.5: Capacity subscriptions

The aspect that distinguishes capacity subscriptions from the previously described capacity mechanisms is that there is no centrally determined estimate of the optimal volume of generation capacity. Rather, each consumer decides how much generation capacity he wishes to have available reliably, and which part of his consumption he is willing to have interrupted during shortages.

Consider for instance an industrial consumer with a 24-hour production line that consumes 10 MW and a second process that only needs to be operated periodically and that consumes 4 MW. The company may opt to purchase only 10 MW worth of capacity subscriptions, which means it may need to interrupt the second process when the fuses are activated.

Consumers purchase capacity subscriptions in a market supplied by the generating companies. Generating companies are only allowed to sell a volume of capacity subscriptions equal to their available capacity. However, as this is a stochastic variable, there always is a probability that they cannot meet their obligations. A generating company with four units of 600 MW could choose to sell only 1800 MW worth of capacity subscriptions. This would allow one unit to be off-line for maintenance. However, there is a possibility that two or even more units are unavailable simultaneously. As in the case of reliability contracts, the generating company needs to weigh this probability and the cost of the penalties against the revenues from the sale of capacity subscriptions.

6.8 Overview

Table 6-1 provides an overview of the capacity mechanisms that were discussed in this chapter.

Table 6-1: Overview of proposed solutions

Method	Description
Capacity payments	Independent agent pays generators for keeping capacity available. In theory, the payments reflect the social value of reliability. Examples: Spain, Argentina (formerly), Columbia, Chile.
Strategic Reserve	An independent agent, usually the system operator, maintains a reserve of power generation units which it dispatches only when the reliability of supply is threatened. The price for electricity should be set high enough not to deter investment. Example: Sweden.
Operating reserves pricing	The system operator purchases operating reserve capacity, possibly more than is needed for short-term operations alone. Extra reserves improve the long-term generation adequacy. By contracting them an incentive is provided to generating companies to create more generation capacity. The price paid for the operating reserves influences the investment incentive and would therefore be high enough.
Capacity requirements (ICAP)	Load-serving entities (e.g. retail companies) are required to contract for a fixed percentage of reserve capacity. The cost of contracting generation capacity is passed on to consumers as part of the electricity price. Examples: PJM, New York Power Pool, New England Pool.
Reliability contracts	An independent agent purchases call options from generators which cover total generation capacity plus a reserve margin. If the market price of electricity rises above the option strike price, the regulator calls the options and receives the difference between the spot price and the strike price. This difference is passed on to consumers. Not tried in practice.
Capacity subscriptions	Consumers buy the right to a certain volume of capacity during peak conditions and allow their peak consumption to be physically limited to this volume during periods of scarcity. The costs of reserve capacity are internalized: each consumer pays for the level of reliability that he desires. Not tried in practice.

7 Evaluation of the capacity mechanisms

Using the goals for capacity mechanisms that Chapter 5 presented, a framework for the evaluation of capacity mechanisms is developed in this chapter to analyze the capacity mechanisms that were described in Chapter 6. Specific attention will be given to the conditions of European markets, most of which have a decentralized structure and have significant exchanges with neighboring systems.

7.1 Introduction

In Chapter 5, reasons were given to doubt that a deregulated market for the generation of electricity will continually provide a socially optimal volume of generation capacity. There is a risk of investment cycles, the severe impact of which could be magnified by strategic behavior of generating companies. The crisis in California demonstrated the high social cost of the recurring supply interruptions that are the consequence of a capacity shortage. Even a small capacity deficit has a severely disruptive impact upon society. The cost of generation capacity is small, by contrast. The investment optimum is asymmetric: it is highly sensitive to a disturbance in the form of under-investment, while the social impact of over-investment is limited. Unfortunately, the rational course of action for private investors is to under-invest as this is the safest course of action from their own perspective. Following the precautionary principle, it is therefore rational to implement a capacity mechanism in order to secure generation adequacy.

An evaluation of the capacity mechanisms that were described in Chapter 6 is provided in this chapter. The goals for capacity mechanisms that were developed in Chapter 5 are developed into a set of criteria in Section 7.2. Specific attention will be given to the robustness of the capacity mechanisms in open, decentralized systems such as in most European countries. The capacity mechanisms are evaluated in Sections 7.3 through 7.8. The findings are summarized in Section 7.9.

As far as the author knows, a similar comparison of the policy options has been made only once before (Doorman, 2000). However, this shorter analysis did not include a

strategic reserve, operating reserves pricing or reliability contracts, perhaps because these options were only presented more recently. Moreover, the criteria used for this comparison were not derived from an analysis of the issue of generation adequacy but were general criteria for the evaluation of policy intervention in markets. As a result, an important difference is that Doorman did not consider the role of market power, which, however, is of key importance. In much of the literature, the approach is to critique a certain market design, for instance an energy-only market or an existing capacity mechanism, and to propose an alternative. A framework for systematic evaluation of the alternative capacity mechanisms has not been developed before, whereas policy makers increasingly face the need to make such a choice.

7.2 Criteria

A set of criteria is proposed in this section for the evaluation of the capacity mechanisms that will be presented in Section 6. Chapter 5 described four goals for capacity mechanisms:

- the provision of adequate incentives for investment in generation capacity,
- compatibility with inter-system trade,
- robustness against the abuse of market power, and
- incentives for improving the price-elasticity of demand.

In addition, three general criteria can be applied to evaluate the merits of a proposed change to the market structure: the change should be effective in reaching its stated goals, it should, naturally, be feasible, and it should contribute to the general goal of economic efficiency. This section will develop these criteria with a focus on the effectiveness with respect to the above four goals for capacity mechanisms.

Investment incentives

The first goal for a capacity mechanism is to ensure an adequate level of generation capacity. The first criterion therefore is effectiveness in stabilizing generation investment at a desired level: how certain is the proposed mechanism to achieve its objective, *in casu* the desired volume of generation capacity? A distinction will be made between the effectiveness of a capacity mechanism in an isolated system, which is a basic performance criterion, and in the special but common case of trade with other electricity systems with different market rules, which will be discussed in the next section. It should be kept in mind that in markets with limited demand participation, a certain probability of service interruptions is economically efficient. (See Section 5.4.) Therefore the effectiveness of a capacity mechanism should not be measured by whether it minimizes the risk of service interruptions but whether it will tend to keep them to the social optimum by providing an optimal volume of generation capacity. Thus, the first criterion is as follows.

Criterion 1: Stabilization of generation investment in an isolated market

The effectiveness of a capacity mechanism can be judged by the degree of certainty with

which it can be expected to result in a socially desirable level of generation capacity. Chapter 5 showed that investment risk is a fundamental cause of the possibility of the development an investment cycle. There are two ways to reduce investment risk. The first is to make the demand for generation capacity explicit, so generating companies can project the need for new capacity more easily and do not need to estimate it from other variables, such as the expected development of electricity prices. The second way to reduce investment risk is by stabilizing the generator revenues. The effectiveness of a capacity mechanism can be judged by the extent to which it performs these two functions.

Compatibility with non-firm imports

Many electricity systems have considerable exchanges with neighboring systems. Ideally, either the same or similar capacity mechanisms should be implemented in interconnected systems. In practice, this may not be feasible, which raises some issues for electricity systems who wish to implement a capacity mechanism before their neighbors do. Therefore an important aspect of the effectiveness of a capacity mechanism is how it performs in the presence of exchanges with other electricity systems that do not have a similar mechanism in place. In principle, connected electricity systems can function as a single market. However, often the connected electricity systems have different rules, for instance because they are at different stages of liberalization or use different models of liberalization. This complicates the inter-system dynamics and the analysis of the security of supply for individual systems.

For systems that normally export electricity, trade is not a threat to generation adequacy. Importing systems, however, may find that trade displaces local investment in generation capacity. This need not be a problem if the imports are as firm as domestic generation capacity. For this purpose, the import contracts would need to be accompanied by firm transmission rights for the same duration. However, many European borders are congested. Explicit auctions, the currently preferred manner to distribute scarce interconnector capacity, typically provide transmission rights for up to a year, which are not necessarily firm rights. (See also Chapter 10.) This poses an obstacle to securing generation capacity with long-term contracts.

If the conclusion is that its future availability is uncertain, the choice may be made to rely only upon generation capacity within the system. In this case, it must be ensured that the capacity mechanism stimulates investment within the system, and not in neighboring systems.

Criterion 2: Effectiveness in securing generation resources in an open market

The compatibility with trade is judged by the extent to which the mechanism's effectiveness is reduced by the presence of a significant volume of imports that are not as firm as electricity produced within the system. This issue is closely related to the next criterion, which concerns the operational aspects of inter-system trade.

Robustness against a regional shortage

When implementing a capacity mechanism in a system with strong interconnections with electricity systems with a different market design, care should also be taken that during a regional shortage (an electricity shortage that does not only affect the system at hand but also its neighbors), the capacity mechanism still is effective. If a capacity mechanism causes the development of an adequate volume of generation capacity but has no means to reserve this capacity for the consumers within the system, they may still have to compete with those from neighboring systems. Consequently, reliability and prices will be the same in all interconnected systems (apart from the consequences of congestion). How can the supply of electricity be secured in the long term, when part of the electricity comes from other systems with energy-only markets (in which there is no system to ensure generation adequacy)? This question is particularly relevant for the implementation of a capacity mechanism in European countries, as the issue of generation adequacy is left to subsidiarity (Art. 7, Directive 2003/54/EC).

Another issue in the EU is that the reliability of imports is reduced by Article 24 of the Directive, which is often interpreted as allowing exporting systems to temporarily halt their exports in an emergency.[38] It is true that this article may only be applied in extreme cases but these are just the cases at hand. If there is no shortage in one of the two interconnected systems, trade should alleviate the shortage in the other system and Article 24 will not need to be applied. If both interconnected systems face a shortage, this article may mean that the system operator of the importing system will need to impose service interruptions, regardless of the contractual arrangements which the market players have made.

Criterion 3: Robustness against a regional shortage

A capacity mechanism is robust against a regional shortage if it makes generation capacity available with priority to the consumers within the system, who, after all, are the ones who paid for it.

Market power in the electricity market

A capacity mechanism should also stimulate generating companies to maximize their output during periods of scarcity, and not to withhold power when it is needed most. (This is the third goal for a capacity mechanism that follows from Chapter 5.) The experience in California and the analysis in Chapter 5 showed that during shortages, in energy-only markets strong incentives may arise to manipulate prices by withholding generation capacity from the market. The probability that a shortage develops can never be ruled out completely but it may be possible to remove the economic incentives for the generating companies to withhold generation capacity.

[38] "In the event of a sudden crisis in the energy market and where the physical safety or security of persons, apparatus or installations or system integrity is threatened, a Member State may temporarily take the necessary safeguard measures. Such measures must cause the least possible disturbance in the functioning of the internal market and must not be wider in scope than is strictly necessary to remedy the sudden difficulties which have arisen." (Art. 24, Directive 2003/54/EC.)

Capacity mechanisms that focus on creating sufficient generating resources should be effective in reducing periods of scarcity to the economic efficient minimum but may not necessarily remove the opportunity for price manipulation (Hobbs et al., 2001c). To keep the inevitable periods of scarcity from lasting longer and costing consumers more than necessary, there should be an incentive to maximize generator output. This issue is significant because capacity withholding adds to the probability of service interruptions and may create large income transfers from consumers to producers.

Criterion 4: Robustness against the abuse of market power in the electricity market

This criterion can be judged by analyzing the short-term incentives that generating companies have during periods of scarcity. Even if the demand is fully price-inelastic, generating companies should still have a positive incentive to maximize their output.

Manipulation of the capacity mechanism

It should be taken care that a change in the market design should not introduce new opportunities for the abuse of market power. Another criterion therefore is to avoid opportunities for manipulation, in particular price gouging through capacity withholding. The experience in the PJM system teaches us that the creation of a new market for generation capacity may offer new opportunities for the abuse of market power. For instance, whenever there is a firm, regulatory demand for generation capacity, it may be possible to manipulate prices by not offering all capacity in this market. Care should be given that a method that mitigates market power in the short-term markets does not create new avenues for market power.

Criterion 5: Robustness against manipulation

A priori evaluation of the robustness of a capacity mechanism against manipulation is difficult, as one can never be sure to have foreseen all the possible games that market players may invent. However, a central aspect to watch for is the effect of withholding generation capacity, whether in the energy market or in a separate capacity market. Crucial is the price-elasticity of demand, as Stoft (2002) remarks for the PJM case: the lower it is, the stronger the incentive to withhold generation capacity. Whereas the price-elasticity of the general demand may not be influenced easily, a number of the capacity mechanisms create an artificial demand for generation capacity, the elasticity of which is impacted by the regulatory design. Care should also be given that other opportunities for manipulation are avoided.

Demand price-elasticity

The last goal for a capacity mechanism is to improve the price-elasticity of demand. Chapter 5 argued that the low price-elasticity of demand is one of the reasons why energy-only markets are unstable. The lack of participation of consumers in the market is the reason why the price mechanism, which arbitrages supply and demand in regular markets, malfunctions in the electricity market. Improved price-elasticity of demand would reduce the need for peaking capacity, and thereby reduce system cost. It would

further contribute to economic efficiency by reducing the risk of random service interruptions, which are intrinsically inefficient.

Improved demand price-elasticity would also result in a higher utilization rate of peaking capacity, so investment risk – which is one of the potential causes of a business cycle – would be lower. In addition, as was mentioned above, it would reduce the incentive to withhold capacity. Because the real price-elasticity of demand has not been revealed fully, it is not certain that improved demand price-elasticity would render the sector immune from a business cycle or remove the incentives for withholding generation capacity but it would constitute an improvement in multiple respects.

The development of storage technology would achieve similar goals but additional incentives do not appear necessary for this purpose. The price difference between peak and off-peak hours already provides a strong reward for the development of storage technology. Incentives for demand to exhibit more price-elastic behavior, on the other hand, depend upon the design of the market and should be taken into account.

There are two ways to stimulate demand price-elasticity. Demand-side management programs reveal some of the hidden demand price-elasticity in markets by offering financial incentives to specific groups of customers in exchange for the right to curtail their consumption periodically. Demand-side management programs are limited to those consumers with a predictable load, lest the consumers would game the system, and usually also to large consumers because of the transaction costs. They are a partial correction for the current lack of possibilities and incentives for consumers to behave in a price-elastic way. More elegant, and potentially more effective, would it be to provide all consumers with efficient incentives regarding their peak consumption through some sort of variable pricing scheme so the full price-elasticity of demand would be revealed. This, however, would require the presence of real-time meters.

Criterion 6: Stimulation of demand price-elasticity

The performance of a capacity mechanism on this criterion can be evaluated by assessing to which extent implementation of the mechanism creates incentives to consumers for shifting their peak consumption to off-peak moments.

Supply-side efficiency

As improved economic efficiency was one of the main motivations for restructuring the electricity sector, any adjustment to the market design should comply with this goal. The previous section already dealt with the main demand-side issue with respect to efficiency, which is to reveal the hidden price-elasticity of demand. On the supply side, an important issue is the ratio between available generation capacity and peak demand: what is the optimal ratio, and how is it to be achieved? Ideally, it should be found through the supply and demand mechanism but Chapter 5 argued that energy-only markets cannot be relied upon for this purpose. If an energy-only market cannot be expected to achieve the optimal volume of generation capacity, a capacity mechanism should also not reduce the degree of competition on the market. Adjustments to the design of the market should not

increase entry barriers or otherwise contribute to the development of an oligopolistic structure.

Criterion 7: Supply-side efficiency

An indicator for the theoretical economic efficiency is to what extent the market decides upon basic parameters, such as the reserve margin, and which ones are determined through a planning process. Given the many uncertainties and the incompleteness of information about the market, it is a safe assumption that a planning process always errs to some extent. Therefore a capacity mechanism will be judged to have a higher economic efficiency the more of its parameters are determined through market forces.

A second aspect is the impact of the capacity mechanism upon the competitiveness of the market. This can be estimated by judging whether the proposed capacity mechanism reduces entry barriers, improves transparency, et cetera.

Feasibility

Naturally, a capacity mechanism should be feasible. The first question is whether, in the terms of the conceptual framework of Figure 3.8, technical changes to the system are required or whether it can be implemented through adjustments of the economic subsystem alone. The second issue is whether these changes are compatible with the existing juridical and institutional structure of the market. Examples of issues to address are whether the proposed capacity mechanism requires an integrated system or whether it can be implemented in a decentral system and whether it requires activities by the regulator or the system operator that exceed their current mandate.

Criterion 8: Feasibility

The feasibility can be judged by the extent to which physical changes to the system are necessary and their cost, by the need for new institutions, and by the degree to which the rules of the system needs to be adjusted.

Compatibility with decentralized systems

There is one last issue to be considered. Some of the capacity mechanisms, such as capacity requirements, have been designed for integrated electricity systems. Our focus is upon European electricity markets, the majority of which are decentralized. In decentralized systems, the system operator has a more limited authority to intervene in the market than in an integrated system. Therefore it is the question as to whether capacity mechanisms that have been developed for integrated systems can be implemented in a decentralized system and, if so, whether they need to be adjusted.

Criterion 9: Compatibility with a decentralized system

In the evaluation of capacity mechanisms, it will only be considered whether the capacity mechanism as it has been implemented or described in the literature is compatible with a decentralized system. Section 8.2 will consider how capacity mechanisms that are not compatible with a decentralized system but that appear promising otherwise can be

adjusted.

Performance with respect to the criteria will be indicated as follows:

--	very poor
-	poor
±	mediocre
+	good
++	very good

Overview

Figure 7.1 shows how the criteria for capacity mechanisms relate to the goals that were formulated in the introduction.

7.3 Capacity payments

Stabilization of generation investment in an isolated system

Borenstein and Holland (2002) provide a fundamental critique of capacity payments. For the special but not entirely unlikely case in which there are no consumers on real-time pricing and the capacity payments are financed through an excise tax on electricity, they show that the combination of the tax and the payments has no net result. The tax would reduce average demand, and therefore the equilibrium level of installed capacity, by the same amount that the payments would increase it. In the more general case that part of consumption is on real-time meters, they derive that the capacity payments will not achieve even the second-best optimum (given the presence of flat-rate consumers).

There are also more pragmatic objections against capacity payments. Capacity payments are intended to stabilize the revenues of generating companies by providing a certain payment per unit of generation capacity.[39] However, they do not make the future demand for generation capacity more explicit than it is in an energy-only market. While the payments improve the average profitability of generation capacity, they do not necessarily make investment in peaking capacity sufficiently attractive. Nor do they provide a clear and timely signal that new investment is demanded. Consequently, the effect of the payments is not clear, which makes it difficult to determine how high they should be (Vázquez et al., 2002).

The uncertainty about the effect of the payments upon the availability of generation capacity is partly caused by the fact that the generation market still may be subject to a business cycle. During periods of ample capacity, the capacity payments may only

[39] Whether the payments really stabilize generator revenues depends among others upon the specifics of how they are calculated. In Argentina and Columbia, some plants were actually exposed to highly volatile payments (Pérez-Arriaga, 2003).

contribute to higher generation profits rather than new capacity, while the generators' lack of information regarding future demand may still cause them to be too late with new construction. It is difficult to create fixed capacity payments that are effective without providing too much subsidy to generators.

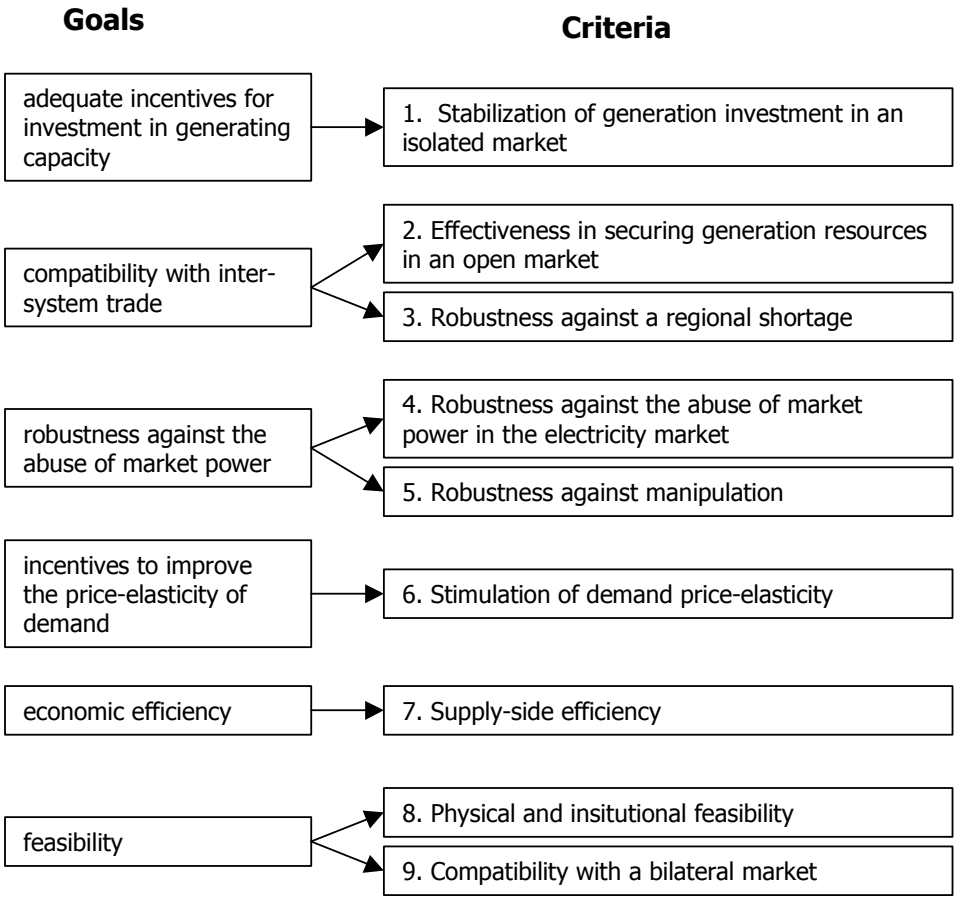


Figure 7.1: Goals and criteria for capacity mechanisms

The Appendix shows that a capacity payment equal to 50% of the fixed costs of a new plant still leads to an investment cycle, even in the static environment of the model, even in the static environment of the model (Figure A.11b). A 75% subsidy is more likely to stabilize investment, as Figure 7.2 shows. This figure shows the model results for the years 2004-2030. On the first Y-axis, peak demand, total generation capacity and investment decisions are plotted. In the model, investment decisions lead to additional generation capacity with a delay of five years. On the second Y-axis are total generator revenues, which are the sum of electricity sales and capacity payments. In the model, their long-run average is normalized (by adjusting the investment response to prices) to the long-run marginal cost of generation, as in a competitive market it is to be expected

that average revenues equal average costs. Despite the absence of shortages in the model, this still is not a robust system; the capacity margin is so limited that even modest fluctuations in the demand growth rate may strongly aggravate the limited cyclical behavior that already is observed.

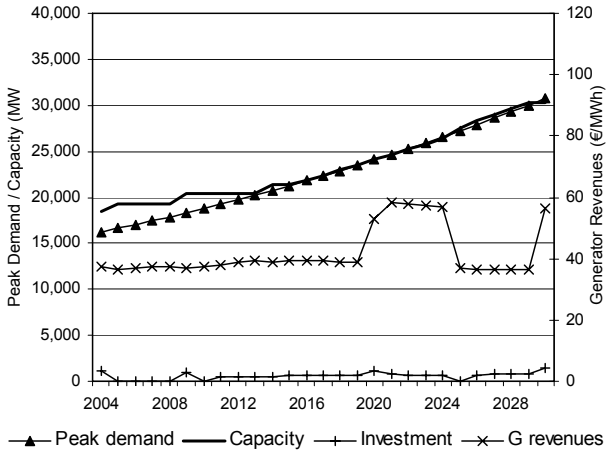


Figure 7.2: The effect of a 75% capacity payment in the model of the Appendix[40]

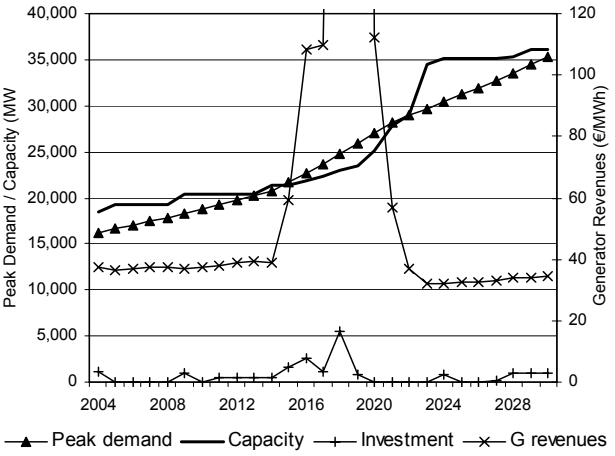


Figure 7.3: Model results of a 75% capacity payment combined with a demand shock of 15% extra growth over 5 years time[41]

[40] This is Figure A.16 in the Appendix.
[41] This is Figure A.29 in the Appendix.

Figure 7.3 gives an example of the impact of a demand shock, where demand grows by an additional 15% over seven years (this is an additional 2 percentage points per year). An example of such a demand shock could be a reduction in the availability of imports, which would increase the demand for domestic production. A demand shock also is a convenient way of modeling a sudden reduction of generation capacity, for instance due to the phasing out of a certain kind of generator technology or due to fuel scarcity. Figure 7.3 represents an extreme scenario but it is a useful test case for robustness of the different capacity mechanisms. In reality, the growth of demand may exhibit cyclical behavior caused by the business cycle of the economy, as a result of which an investment cycle would develop much sooner in the electricity market than in the model. This was demonstrated in Figure 5.10 on page 95.

Dynamic capacity payments are intended to solve the question of how to establish the correct level of the payments, as the payment level increases with the need for capacity. However, the fact that they fluctuate on a very short-term basis (e.g. hourly) presents a new issue, as the fluctuating nature of the incentive comprises a risk to investors. This will cause investors to depreciate the value of future payments, which reduces their effectiveness in providing stable, long-term incentives. There is a discrepancy between the short-term nature of the price signal and its intended long-term effect (Jaffe and Felder, 1996).

Effectiveness in securing generation resources in an open market

If applied to domestic generation capacity alone, capacity payments would make domestic electricity cheaper than imports, *ceteris paribus*. However, in the presence of significantly cheaper imports, the capacity payments would need to be proportionately higher if the goal is to maintain generation adequacy through generation capacity within the system.

Robustness against a regional shortage

During a regional shortage capacity payments provide no means to keep electricity from being exported. Therefore the price spikes will be as high as in neighboring systems and the reliability will also be the same. The exception is that if the payments lead to a sufficient volume of generation capacity within the system, EU countries appear to have the option to temporarily curtail exports during a crisis (Art. 24, Directive 2003/54/EC).

Robustness against market power in the electricity market

A disadvantage of payments for installed capacity is that they do not reduce the incentive to withhold capacity during shortages. The payments are made for generator's potential to provide power, not for their actual contribution to maximizing reliability (Vázquez et al., 2002). To mitigate this incentive, the capacity payments can be accompanied by an electricity price cap. This would also, however, reduce incentives for demand to exhibit price-elastic behavior. More fundamentally, the reduction in generator revenues due to the price cap would need to be known in order to calculate the correct level of capacity payments. It is precisely the problem of energy-only markets that it is difficult to determine the expected revenues from price spikes. If generating companies are not able

to estimate them due to a lack of information, it is highly questionable that a central planner would have better information and be more successful. A solution could be to make payments for available capacity but they may also be manipulated. Payments for available capacity may cause unavailable capacity to be declared available; the owners could prevent the capacity from being called upon by bidding very high in the energy market.

Robustness against manipulation

The dynamic capacity payments in the former England and Wales Pool appear to have been manipulated, which resulted in overpayments to the generating companies (Wolak and Patrick, 1997; Von der Fehr and Harbord, 1998).

Stimulation of demand price-elasticity

Theoretically, optimal capacity payments should not alter the incentives for demand compared to an energy-only market, except if they are combined with a price cap to mitigate market power.

Supply-side efficiency

In a static equilibrium, the payments should compensate for an insufficient incentive to invest. The payments would shift the equilibrium volume of generation capacity to a higher level; if the payments would be set correctly, they could shift capacity to the social optimum. Therefore the payments would be efficient in theory. The payments would need to be accompanied by a price cap equal to the VOLL to protect consumers against overcharging, similar as in an energy-only market, as was mentioned in Section 5.2.

The problem is that the relationship between the level of payments and the investment response is not known, so it is difficult to determine the optimal level of payments. If, for the purpose of mitigating market power, a choice is made for a combination of higher capacity payments with a lower price cap, there is the additional difficulty of estimating the lost profits due to the price cap and due to the loss of efficiency due to a reduction in demand response.

Feasibility

Fixed capacity payments are easy to implement but there may be legal obstacles. At least within the EU, they may be considered as state aid, and therefore against the competition rules. This is not, of course, how they are intended, as they should be cost-neutral to consumers and revenue-neutral to generating companies but the legal interpretation may be different. Moreover, an electricity system with congestion on its boundaries would probably choose to offer the payments only to generators within its limits, which could be considered discriminatory and a distortion of the ‘level playing field’.

Dynamic capacity payments can only work in integrated systems, while the preferred model in Europe is that of a decentralized market. In decentralized markets it may be difficult to obtain the exact data regarding the availability of generation capacity on a continuous basis, which would be necessary to calculate the payments.

Compatibility with a decentralized system

There is no obstacle to implementing capacity payments in a decentralized system.

Conclusion

Capacity payments have fundamental shortcomings. Even in an isolated system, it is questionable whether they sufficiently stabilize generation capacity. In an open system, they provide no means to guarantee that the electricity always is available to those who have made the capacity payments. Finally, capacity payments do not change the electricity market dynamics, so capacity withholding may still be lucrative. This effect can be mitigated through a price cap but the correct level of this price cap is difficult to establish. Moreover, it may be circumvented by exporting and re-importing the electricity. An advantage of capacity payments is the simplicity of the system, which makes it transparent and easy to implement. Table 7-1 presents a summary of the performance of capacity payments.

Table 7-1: Evaluation of capacity payments

Stabilization of investment	-
Effectiveness in securing generation in an open market	+
Robustness against a regional shortage	-
Robustness against market power in the electricity market	±
Robustness against manipulation	±
Stimulation of demand price-elasticity	±
Supply-side efficiency	±
Feasibility	++
Compatibility with a decentralized system	+

7.4 Strategic reserve

Stabilization of generation investment in an isolated system

At first glance, a strategic reserve appears a robust way of securing generation adequacy, as the agent who manages the system can purchase as much reserve capacity as he wishes. This, however, does not change the overall volume of generation capacity in the short term. The long-term effectiveness depends on the investment signal that is sent to the generation market. The market must be inspired to replace at least part of the generation capacity that the reserve withdraws from the market. To encourage investment, the *de facto* price cap which is created by the price at which the reserve is dispatched, P_{sr}, must be high enough.

If P_{sr} is set at the average value of lost load, the dynamics of the system are not changed, relative to an energy-only market, with the exception that the probability of service interruptions is reduced by the presence of the reserve. In theory, there would be no need for a reserve if prices are allowed to reach the average value of lost load, so the size of the reserve depends entirely on the degree to which the market fails to reach the optimal volume of generation capacity. If P_{sr} is set equal to the average value of lost load, the tendency towards investment cycles will not be dampened, nor will the frequency and duration of price spikes be altered. The benefit of a reserve is limited to a reduction of service interruptions. However, as prices rise to the average value of lost load, consumers should on average be indifferent about this reduction.

At a lower price cap, the reserve will need to play a more active role in the market. The reserve will take over the provision of a part of peak demand. As in a system of operating reserves pricing, prices will spike more often than in an energy-only market but the height of the price spikes will be limited by the availability of the strategic reserve capacity at a price P_{sr}. This may have a dampening effect upon investment cycles because the frequency and height of the price spikes, and hence expected generator revenues, are more predictable. There still is no clear signal to the market, however, regarding the optimal volume of installed capacity, and while generator revenues will be more stable than in an energy-only market, they may still be difficult to forecast. Therefore a strategic reserve may still not provide sufficient investment incentive to avoid investment cycles.

The difficulty is to establish the relationship between the volume of capacity that is held in reserve and the accompanying optimal price cap P_{sr}. Example 6.1 showed how much detailed data is required to make a correct estimate of P_{sr}; errors would either undermine the investment signal or create undue income transfers from consumers to generating companies. The price cap determines the income of the marginal generator in the market. To calculate how large a volume of generation capacity will be profitable for the generating companies, the agency that establishes the price cap needs to know the price-duration curve, including its stochastic distribution. This information is not sufficiently available in many markets, due to their short history, which introduces a risk that the reserve is dispatched at an inefficient price. This would either lead to too low an incentive to invest in generation capacity, or to too high a volume of reserve capacity.

The operator of the strategic reserve makes use of the fact that he has market power when the reserve is needed, so he can determine the reserve price. In the short run, this reduces consumer welfare, as the prices are above the competitive level. This is necessary, however, to attract investment. Were the reserve offered at a price equal to the marginal cost of generation, we would be back at an energy-only market, with the role of the reserve operator no different from other suppliers of generation capacity. However, the fact that consumer welfare can be improved in the short run by lowering the reserve dispatch price means that there may be considerable political pressure to do so during a period of scarcity. This – even the threat of this – would undermine the incentives for investment.

A strategic reserve may be useful when generating companies are about to decommission or mothball generating units while generation adequacy is not expected to be maintained

in the future. Then the old units can be purchased and kept available. Due to the many difficulties, however, this should only be considered as a short-term measure while better investment incentives are being developed.

Effectiveness in securing generation resources in an open market

Imports that are not firm for at least several years would directly undermine the effectiveness of a strategic reserve in guaranteeing generation adequacy. These imports displace local generation capacity but do not provide any certainty about their future availability. To maintain generation adequacy, the extent to which imports displace local generation capacity would need to be compensated with extra reserve capacity locally. The Netherlands, for instance, has import capacity on the order of 20% of peak demand. If domestic generation capacity is displaced by this amount, a strategic reserve of at least the same volume would be needed to achieve nominal self-sufficiency, without even a capacity margin. The operator of the reserve only has an indirect way to control the volume of generation capacity within the system: by increasing the reserve size, he hopes that more investment will take place within the system. As long as there is sufficient interconnector capacity available, however, the investment may take place elsewhere. If the future availability of this capacity is uncertain, this has a limited effect upon reliability. Therefore this method does not provide a firm means to secure generation adequacy within an open system.

Robustness against a regional shortage

In an open market, the system operator may control the output of the strategic reserve, but the remainder of the generation capacity is free to be sold to the highest bidder, including those in neighboring countries. This means that, despite the presence of the reserve, ultimately system reliability is no better than that in neighboring systems. Moreover, the market prices in the interconnected systems would also converge, regardless of P_{sr}, so the presence of the reserve would not limit the height of regional price spikes. It must be concluded, therefore, that in the presence of significant exchanges with neighboring energy-only markets, the effectiveness of a strategic reserve in enhancing reliability and stabilizing prices is severely limited (even if it would be successful in securing an adequate volume of generation capacity within the system).

Robustness against market power in the electricity market

Generating companies have an incentive to withhold generation capacity to raise prices, the same as in an energy-only market, until the electricity price reaches P_{sr}, the level at which the strategic reserve is deployed. Therefore a large strategic reserve with a low price cap reduces, but does not eliminate, the incentive to withhold. The higher P_{sr}, the stronger the incentive to withhold generation capacity. A reserve that is dispatched at $P_{sr}=VOLL$ provides no protection against price manipulation (apart from reducing the risk of service interruptions).

Robustness against manipulation

Generating companies may be able to manipulate the prices at which generators are sold

to the reserve.

Stimulation of demand price-elasticity

A small reserve, priced at the average value of lost load, would not alter the incentives for demand compared to an energy-only market. However, while theoretically optimal, practical market conditions keep many consumers from behaving in a price-elastic manner. See also Section 6.2. A larger reserve with a lower price cap would reduce the incentives for demand to shift to off-peak hours. On the other hand, interruptible contracts could be used as part of the reserve, which would enhance the price-elasticity of demand.

Supply-side efficiency

It may not be possible to create a reserve consisting of only the generating units with the highest marginal costs (presumably the oldest units). If the operator of the strategic reserve cannot purchase a sufficient number of old units to fill his reserve, he may be forced to construct new plant for the reserve. This would probably disturb the economic merit order of dispatch. This risk increases with the size of the reserve.

A second efficiency question arises when a choice is made for a larger, more active strategic reserve. In this case, an independent, supposedly neutral (government) agent becomes an active market participant. In fact, the agent would take over part of the peaking capacity market, eliminating competition in this market segment. This runs counter to the intention of liberalization, which was to increase efficiency through the introduction of competition.

As was mentioned above, a difficulty with designing a strategic reserve (as well as operating reserves, see the next section) is that the system operator needs to know the price-duration curve to be able to determine the relationship between the size of the operating reserve and the correct price at which to dispatch the operating reserve.

Feasibility

From an institutional perspective, a strategic reserve requires a fundamental change for many systems, as a regulated agent – probably the system operator –becomes an active participant in the generation market. This may conflict with rules regarding the unbundling of networks and generation, if the system operator is also the transmission operator. The practical implementation is limited to this agent purchasing the necessary generation units and obtaining the expertise to operate them.

There is a question as to whether the high price that should be charged for electricity from a strategic reserve is politically acceptable. It may be difficult to explain why old generators, which have been paid for, would need to be so expensive (reflecting scarcity, rather than marginal costs). The risk that during a prolonged power crisis political pressure would force the price of the strategic reserves to be lowered would constitute an investment risk for generating companies and therefore discourage investment. It appears that this method creates its own regulatory uncertainty, at least until the reserves have

actually been used during a shortage.

Compatibility with a decentralized system

There is no obstacle to implementing capacity payments in a decentralized system.

Conclusion

A strategic reserve may dampen investment cycles but will probably not eliminate them. The mechanism is not very robust against demand shocks, nor is it effective in an open market. It is difficult to choose the parameters of the capacity mechanism correctly, so there is a risk that the incentive for private generating companies to invest is reduced too much.

A strategic reserve that is dispatched at a price equal to the average value of lost load (P_{sr}=VOLL) does not alter the dynamics of the market, except that it reduces the risk of service interruptions – but at a price at which the average consumer is indifferent. The probability of investment cycles, with the accompanying prolonged periods of high prices, and the opportunities for price manipulation through capacity withholding, remain unmitigated.

A strategic reserve that is dispatched at a lower price (P_{sr}<<VOLL) limits but does not eliminate the function of price spikes to signal investment. As a result, there remains a possibility of the abuse of market power during periods of scarcity (by withholding power until the strategic reserve is deployed) and of the development of investment cycles (albeit less than if P_{sr}=VOLL). Table 7-2 summarizes the performance of a strategic reserve on the various criteria for these two situations (P_{sr}=VOLL and P_{sr}<<VOLL).

Table 7-2: Evaluation of strategic reserves

	P_{sr}=VOLL	P_{sr}<<VOLL
Stabilization of investment	--	±
Effectiveness in securing generation in an open market	-	-
Robustness against a regional shortage	-	-
Robustness against market power in the electricity market	--	±
Robustness against manipulation	+	+
Stimulation of demand price-elasticity	+	±
Supply-side efficiency	-	--
Feasibility	±	++
Compatibility with a decentralized system	+	+

The EU's tendering procedure

Section 6.3 mentioned that the EU's tendering procedure appears to resemble either a strategic reserve or capacity payments (Art.7, Directive 2003/54/EC). As such, it would have the same advantages and disadvantages as these mechanisms. An added disadvantage of the procedure is the long time trajectory that needs to be followed: first the need for new capacity needs to be demonstrated through monitoring of the market, then there is a half year's notice for the tender, then the tendering procedure needs to be executed, and finally of course the generation capacity needs to be constructed. Chapter 5 described the risk of investment cycles, due to the long lead time for new generation capacity. The tendering procedure would not change this; to the contrary, the tendering procedure itself adds at least half a year to the lead time for new capacity.

Questions remain about the tendering procedure, such as how it will be financed and, especially, who will control the generation units built under this procedure and under which conditions (and against which price) they will be operated. It leaves the possibility open of providing capacity payments only to new generators, which would distort the market and discourage any new investment other than through a new round of tenders

7.5 Operating reserves pricing

Stabilization of generation investment in an isolated system

Operating reserves pricing can be considered a mitigated form of an energy-only market (Stoft, 2002). The incentive for investing in peaking generation still comes from periodically occurring price spikes but the price spikes are lower and more frequent, so the investment signal is more stable and predictable than in an energy-only market. Consequently, the potential for the development of a business cycle is reduced but not removed. The system operator may attempt to counter a business cycle by adjusting the volume of the operating reserve in a counter-cyclical manner but he has no other instrument for intervention if he considers the current volume of available generation capacity too low.

Figure 7.4 shows the effect of operating reserves pricing should have upon the market: price spikes occur more frequently but are limited and generally are not accompanied by outages. In the idealized environment of the model, the investment cycles dampen out and generation capacity develops in a perfectly stable manner. In practice, changes in the growth rate of demand will continuously shift the market equilibrium, as a result of which investment cycles probably will continue to exist. Their magnitude will generally be lower than in an energy-only market, and as the average capacity margin will be larger, the probability of outages should also be much lower.

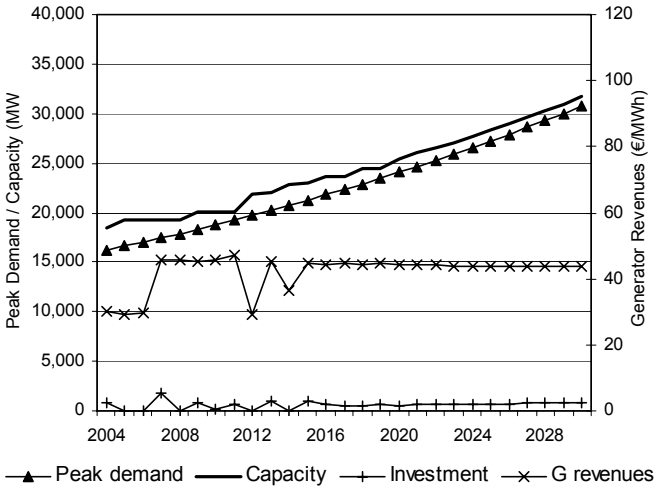


Figure 7.4: Operating reserves pricing may dampen investment cycles (in the model of Appendix A[42])

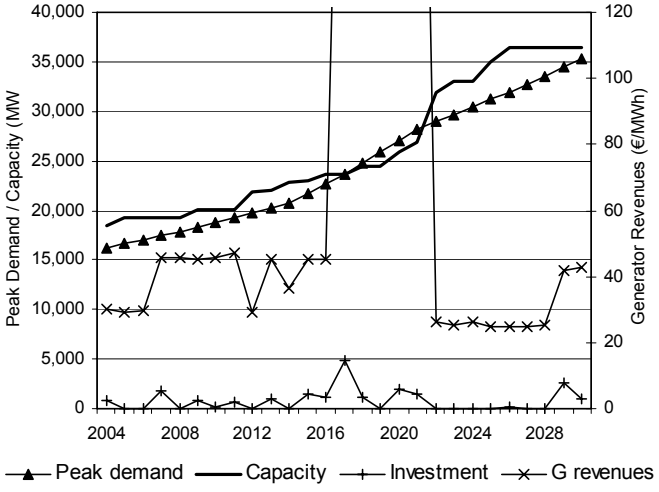


Figure 7.5: Model results of operating reserves pricing with a demand shock[43]

While operating reserves pricing should lead to a larger capacity margin than an energy-only market and therefore be more stable, changes in the growth rate of demand may still cause shortages if the combination of the size of the reserve and the strength of the

[42] This is Figure A.17 in the Appendix.
[43] This is Figure A.30 in the Appendix.

investment response to price signals is insufficient. When supply becomes limited in a market with operating reserves pricing, the average electricity price during that period may be higher than in an energy-only market. This can be seen as follows In an operating reserves pricing market, as in an energy-only market, the electricity price should (in theory) rise to the average value of lost load when supply is insufficient to meet demand. Presumably, these occasions will be fewer than in an energy-only market that is subject to an investment cycle, but considering that there still should be some hours of load shedding if the volume of available generation capacity is optimal, this possibility continues to exist. So during the real peaks, the price will be the same as in an energy-only market. During the shoulders, however, the price will drop less quickly than in an energy-only market, as the electricity price will be influenced by the system operator's demand for reserve capacity and therefore stay at P_{or}, while in an energy-only market competition should keep the electricity down to the marginal cost of generation or, be it the case, the cost of interruptible contracts. When there is no acute shortage of generation capacity, this mechanism creates a timely investment signal. During an episode of scarcity, however, it prolongs the price spike substantially, leading to a significantly higher average electricity price. See Figure 7.5. Appendix A describes this phenomenon in more detail.

Example 6.2 on page 118 showed that to choose the correct P_{or}, given a certain volume of operating reserves, the following information is required:

- the fixed costs of the marginal generator
- the average value of lost load
- the load-duration curve

The later two may be difficult to obtain. Section 5.2.4 and 5.3.1 mentioned the difficulties with estimating the average value of lost load. In principle, obtaining the load-duration curve simply is a matter of measuring electricity consumption carefully. In practice, this may not be done (like in the Netherlands), the information may not be publicly available, or the available time series may not be long enough to make an accurate estimate of the stochastic distribution of the curve.

The design of the system is sensitive to a correct estimate of these data. If P_{sr} is established at a wrong level, the total volume of generation capacity (the sum of the capacity in the operating reserve plus the capacity in the market) would be sub-optimal. Moreover, If P_{sr} is set too high, it would also create more opportunities for price manipulation. The actual effectiveness of the system, in terms of stabilizing investment, depends upon whether the generating companies can estimate the resulting price spikes accurately, for which they need the same information.

Effectiveness in securing generation resources in an open market

An operating reserve has the same vulnerability against non-firm imports as a strategic reserve. Operating reserves pricing influences the volume of generation capacity within the system indirectly. In principle, operating reserves pricing leads to limited price spikes, which occur more frequently than the price spikes in an energy-only market. If a system with operating reserves pricing is linked to an energy-only market, the higher

prices in the market with operating reserves pricing will attract imports from the energy-only market. As a result, the investment signal will partly 'leak' to the neighboring system.

Robustness against a regional shortage

The presence of sufficient generation capacity to meet peak demand provides no guarantee that the reliability will be higher than in neighboring systems. During a regional electricity shortage, imports may only be available at the market price of the neighboring systems, which can be expected to approach the average value of lost load. As a result, both the electricity price and the reliability will be the same as in the neighboring systems (The Brattle Group, 2003). Arbitrage between open markets means that operating reserves pricing is not effective if the neighboring systems do not implement a similar capacity mechanism. Only when exports are limited by network congestion may the reliability within the system with operating reserves pricing be higher than the reliability of neighboring systems.

Robustness against market power in the electricity market

An advantage of operating reserves pricing, Stoft (2002) argues, is that it reduces the incentive for withholding capacity. In its simplest form, operating reserves pricing creates a *de facto* price cap (determined by the system operator's purchasing price P_{or}) which may be much lower than the average value of lost load. The lower P_{or}, the lower the incentive to withhold power. As a result, the potential for profits from withholding is greatly reduced. In addition, the price spikes that occur in a system of operating reserves pricing are not only lower but longer. Withholding may cause prices to rise to the price cap sooner than in a competitive situation. However, with lower and longer price spikes, the relative impact upon total generator revenues is smaller than in an energy-only market. Once the market price has reached the price cap, there is no more incentive to withhold generation capacity. This effect is similar to the mitigation of market power achieved by a strategic reserve, with P_{or} equivalent to P_{sr}.

As Section 6.4 already mentioned, Stoft (2002) actually argues that the system operator should not use a single price P_{or} but a downward-sloping demand function, in order to reduce market power. This means that an increase in the market price will cause the system operator to purchase a smaller volume of reserves. This would further reduce the attractiveness of capacity withholding, as its effect upon the market price would be lessened. Determining this demand function so it provides an optimal balance between investment incentives and market power mitigation may be quite difficult.

Robustness against manipulation

As with every change in the market design, caution must be given to the potential development of new opportunities for strategic behavior. In particular, the actual availability of capacity that is sold as operating reserves must be verified. One way of doing this is to occasionally dispatch it out of merit order.

Stimulation of demand price-elasticity

The lower price spikes reduce the incentives to demand for limiting peak consumption. In this respect the theoretical efficiency is lower than that of an energy-only market. A trade-off is to be made between a high price cap, which does little to reduce the risk of investment cycles (although the reserve provides a buffer that reduces the risk of service interruptions) and the incentive to withhold capacity during periods of scarcity, and a low price cap, which distorts demand-side incentives. On the other hand it is possible to allow interruptible contracts for at least part of the reserve requirements, which adds an incentive to develop these contracts and restores at least part of the demand-responsiveness.

Supply-side efficiency

As long as the price of the operating reserve P_{or} exceeds the variable cost of production of the most expensive generating unit in the system, operating reserves pricing should not reduce the efficiency of the generation market. It avoids the inefficiency of distorting the merit order dispatch, which a strategic reserve may do, by contracting the reserves from the market because market parties would offer the units with the highest operating costs to be contracted as reserves. As with a strategic reserve, an obstacle is the necessary information (in particular the load-duration curve and its stochastic distribution) to determine the optimal combination of reserve volume and willingness to pay P_{or}.

Feasibility

A significant advantage of operating reserves pricing is that it is easy to implement, as it is an expansion of an existing activity by the system operator. Every system operator needs operating reserves: this system simply requires that the system operator pays for their availability.[44] If desired, the system operator contracts more capacity than he would need for maintaining operational stability of the system alone, in order to stabilize investment in generation capacity more strongly.

If P_{or}=*VOLL*, the same question of political acceptability may arise as for a strategic reserve: if the system operator has contracted reserves, why would they need to be dispatched at such an extreme price? While operating reserves pricing is similar to maintaining a strategic reserve in many ways, it creates less of a conflict with the principle of unbundling because the system operator does not own or dispatch the reserves (other than for operational stability, which the system operator already does). To the contrary, when scarcity develops, the system operator purchases less generation capacity. Therefore it may be easier to implement in systems where the principle of unbundling of network operation and the generation market is legally enforced.

Compatibility with a decentralized system

There is no obstacle to implementing operating reserves pricing in a decentralized

[44] This is not automatically the case. In the Netherlands, for instance, generating companies are required to offer unused capacity to the balancing market, but are only remunerated if they are called, except for a few long-term contracts for operating reserves.

system.

Conclusion

Operating reserves pricing is somewhat similar to maintaining a strategic reserve. Like the dispatch price of a strategic reserve functions as a *de facto* price cap in the market, the willingness to pay for operating reserves limits the market price of electricity. Thus, again there is a choice between a small reserve with a high price cap and a larger reserve with a correspondingly lower electricity price cap. The higher the system operator's demand for operating reserves (and the lower the corresponding prices), the smaller the potential damage from capacity withholding.

Table 7-3: Evaluation of operating reserves pricing

	P_{or}=VOLL	P_{or}<<VOLL
Stabilization of investment	--	±
Effectiveness in securing generation in an open market	-	-
Robustness against a regional shortage	-	-
Robustness against market power in the electricity market	--	±
Robustness against manipulation	+	+
Stimulation of demand price-elasticity	+	±
Supply-side efficiency	±	±
Feasibility	±	++
Compatibility with a decentralized system	+	+

Operating reserves pricing suffers from some of the same fundamental weaknesses as a strategic reserve. Its effectiveness in dampening investment cycles is limited and is even further reduced in open systems that are interconnected with energy-only markets. During a regional shortage the electricity prices will be set by the interconnected system as a whole (and may therefore be very high) and the reliability of service will also be equal to that of the neighboring systems. The information necessary for the design of an operating reserve may be difficult to obtain.

The main advantage of an operating reserves pricing is that implementation should be simple because this method is a mere expansion of the existing operating reserve. Table 7-3 shows the performance of operating reserves pricing along the criteria.

7.6 Capacity requirements

Stabilization of generation investment in an isolated system

A system of capacity requirements appears to be a simple, effective and efficient way to achieve a desired level of reliability. It establishes a capacity margin much in the same way as prior to liberalization but leaves the provision of all generation capacity to the market. By regulating the total volume of generation capacity, rather than only the size of the reserve margin, this method is more effective than the previously discussed ones.

Figure 7.6 shows how capacity requirements stabilize investment (in the ideal setting of the model in the Appendix). The generator revenues in the figure are the sum of the annual average electricity price plus the annual payments made by load-serving entities for capacity credits.

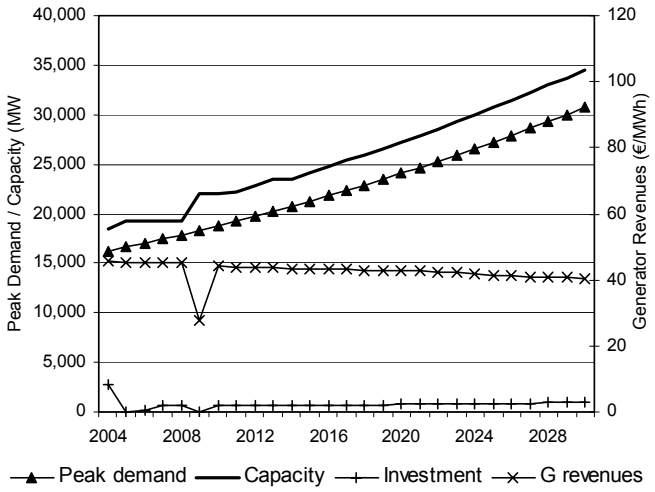


Figure 7.6: Effect of installed capacity requirements upon investment[45]

At the start of the model, the system is short of its reserve requirements, which prompts a small overreaction. As a result, there is excess generation capacity (beyond the capacity requirement) in 2009, which causes the capacity prices to drop to almost zero. The rest of the time, the conservative nature in which the investment decisions are modeled causes investment to lag behind the capacity requirements, so the capacity price always is set by the penalty level. In reality, changes in the growth rate of demand are likely to cause some oscillations in the system, so the capacity price will drop to the marginal cost of providing generation capacity from time to time.

The strong investment signal makes a capacity requirement robust against demand shocks. Figure 7.7 shows that even an additional demand growth of 15% between 2015

[45] Figure A.23 in the Appendix.

and 2022 does not cause the system to become unstable. Prices remain near the long-run marginal cost of generation.

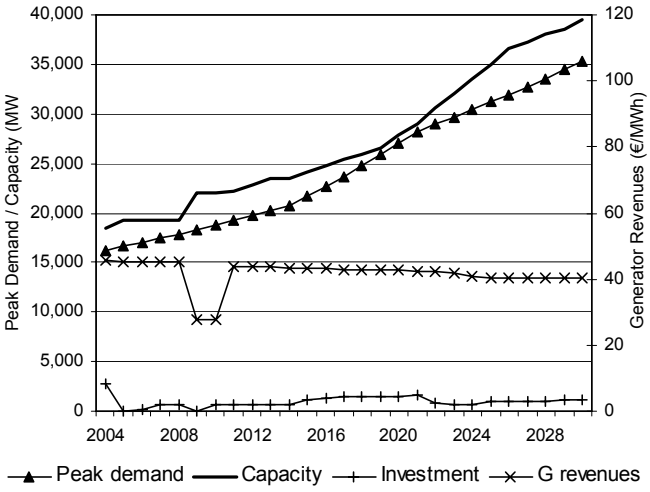


Figure 7.7: Capacity requirement, demand shock[46]

Effectiveness in securing generation resources in an open market

A system of capacity requirements provides a direct means of controlling the volume of generation capacity within the system because the capacity credits can be required to be backed by generation capacity within the system. The issue remains that self-reliance may be a costly solution.

Robustness against a regional shortage

It may be difficult to prevent electricity from being sold to neighboring energy-only markets during a regional shortage. PJM has experienced problems of this nature (Stoft, 2000). However, a pool-based system may offer means to control the destination of electricity, for instance by requiring generators who have sold capacity credits to sell their electricity to the pool (when the generation capacity underlying the credits is recalled), allowing out-of-system consumers only to purchase any excess electricity. Thus, in integrated systems it may be possible to maintain a price cap below that of neighboring systems. However, a shortage in neighboring systems would still result in exports that would push the price up to the cap, so it still would have a welfare impact upon the system.

Robustness against market power in the electricity market

While the presence of reserve capacity reduces the probability of price spikes, once they

[46] Figure A.31 in the Appendix.

occur, there still is an incentive to manipulate prices through capacity withholding. The incentive to withhold can be limited, however, by a maximum price, as the price spikes are not needed to signal new investment (Hobbs et at., 2001b). This is allowable because there is an alternate revenue source for investment in generation capacity. A price cap may limit the exercise of market power and limits the income transfers during shortages.

Another solution, proposed by Shuttleworth et al. (2002), is to require the load-serving entities to purchase capacity in a way that provides them with access to electricity. In fact, they suggest that the load-serving entities are required to purchase options to cover their expected demand plus a margin. This would limit electricity price spikes to the option strike price. This solution is a hybrid of PJM's form of capacity requirements and reliability contracts. It will be further discussed in Section 8.2.

Robustness against manipulation

One of the main difficulties experienced in PJM is that the system can be gamed by providing reserve capacity that is not actually operational: it rewards 'iron in the ground'. In PJM the penalty for unavailability of generators that have sold capacity credits is about equal to the cost of new capacity (PJM Interconnection, L.L.C., 2003). Apparently this is too low, given the probability to be caught, so the expected revenues from selling capacity credits exceed the expected amount of penalties to be paid. Theoretically, this penalty should be equal to the social costs of not being able to produce. If the system is short of electricity in real-time, this means that the penalty should equal the average value of lost load. If there is excess capacity, on the other hand, there are no social costs and the penalty may be set to zero (Shuttleworth et al., 2002). However, capacity is only recalled if the reserve margin is low, so an argument can be made that the penalty simply should equal the average value of lost load.

A second risk, which was mentioned in Section 6.5, is that a firm reserve requirement creates a perfectly inelastic demand for reserve capacity. Not only does this increase investment risk, it also provides a venue for the exercise of market power (Stoft, 2002). The same solution applies as for the problem of the price volatility of the capacity credits. The penalty for non-compliance should be elastic: it should increase with the magnitude by which a load-serving entity does not meet its capacity requirement. This reduces both the volatility of the capacity credit prices and the incentive to withhold generation capacity.

Another practical problem in the initial PJM design was that generators could 'delist' their capacity on short notice (Hobbs et al., 2001b). Thus they could earn revenues in the capacity market when electricity demand was low, and sell at high prices in the (neighboring) electricity market when that was more profitable. The solution was to increase the minimum duration for which a capacity credit may be sold, so generators need to decide for a whole season at once whether to offer capacity credits, and to require a longer notice for de-listing reserve capacity.

Stimulation of demand-side price elasticity

A system of capacity requirements stimulates active involvement of demand, as interruptible contracts can be used for reserve capacity. The stimulation of demand-side management is limited to large consumers with predictable loads, however. The transaction costs of interruptible contracts with small consumers are too high, while consumers with unpredictable loads could game the system too easily. The price-elasticity of demand is limited by the price cap, so it necessarily always is less than in an energy-only market with real-time pricing. If the price cap is high, however, it is questionable whether much demand response is lost.

Supply-side efficiency

Overall economic efficiency of the system depends upon the accuracy with which the system planners can estimate the socially optimal level of reserve capacity and upon the efficiency of the capacity market. An advantage over the previous systems is that only the volume of the reserve is determined through a planning process: the market then establishes the price for reserve capacity. As a result, the difficulty of determining the optimal relationship between price and volume is avoided.

Feasibility

Implementation of capacity requirements requires a substantial amount of administration, as supply and demand must be monitored and each load-serving entities share of the reserve margin must be calculated. A secondary market for capacity credits must be established, including a system to track the credits and verify that they are actually covered by available generation capacity. Finally, the compliance of the load-serving entities must be monitored and enforced.

Compatibility with a decentralized system

Experience with capacity requirements only exists in integrated systems. In the absence of trade with other electricity systems, it appears that the capacity mechanism can be implemented decentralized systems as well, as long as all the required data regarding demand and generator availability are available. The problem is not the implementation in a closed system but the effectiveness in the presence of trade with different markets.

The current design requires generating companies who have sold capacity credits and whose output is recalled to sell to the pool. In a decentralized system, this could be interpreted as being required to offer any capacity that is not covered by long-term contracts to the spot market. This, however, offers no means to prevent the electricity from being sold outside the system during a regional shortage. Absent a method to ensure that capacity that has sold capacity credits is actually applied to in-system consumption, it must therefore be concluded that capacity requirements appear ineffective in a decentralized system with substantial exchanges with markets with a higher price cap. Section 8.2.2 further discusses options for implementation of a system of capacity requirements in a decentralized, open market.

Conclusion

Capacity requirements have important advantages. Foremost is that they are robust, as the total volume of generation capacity within the system is regulated, and they have been proven to be feasible. The experience with this system in the PJM system, and in similar systems in New York State and in the New England Power Pool, provides valuable support for its implementation elsewhere. These are all integrated systems, however; implementation in decentralized systems raises some issues, such as the question how to recall capacity and keep it from being exported during a regional shortage. These issues will be discussed in Section 8.2.

A system of capacity requirements provides incentives for demand-side-management, as producers have the option to meet their capacity requirements through interruptible contracts. It does have some weaknesses, however, in that both the capacity market and the electricity market may be manipulated by strategically withholding generation capacity. Table 7-4 summarizes the advantages and disadvantages of capacity requirements.

Table 7-4: Evaluation of a capacity requirements

Stabilization of investment	++
Effectiveness in securing generation in an open market	+
Robustness against a regional shortage	±
Robustness against market power in the electricity market	±
Robustness against manipulation	±
Stimulation of demand price-elasticity	+
Supply-side efficiency	+
Feasibility	+
Compatibility with a decentralized system	±

7.7 Reliability contracts

Stabilization of investment in an isolated system

The volume of options that the system operator demands provides the market with a clear indication of the demand for available generation capacity. Investment risk is also lowered, as generators convert part of the volatile income from price spikes into a continuous income by selling the call options. Generators' expected total revenues should not change (if they bid well) but their income volatility is greatly diminished. Thus, generators face a stable and clear investment signal, like in a system with capacity

requirements. The effectiveness in dampening investment cycles therefore is judged to be similar to capacity requirements.

Effectiveness in securing generation resources in an open market

With respect to imports, there are two possible approaches. One is that the system operator only purchases call options from domestic generators. This is the best way to secure generation adequacy in an environment of connected electricity systems with different market rules and/or non-firm imports. If all options are purchased domestically, that would stimulate the market to maintain sufficient generation capacity to always meet domestic demand. As a result, structural imports would not lead to a lower volume of installed generation capacity. However, this would create the same inefficiency as was found to be the case with the other capacity mechanisms: to maintain generation adequacy, the full volume of imports would need to be covered by domestic reserve capacity.

It is also possible to treat imports like generation. In this case, importing parties would need to sign reliability contracts, providing them with a strong incentive to make their imports available reliably. Unlike generators, importing parties may not have full control of the availability of the electricity they import, as network operators may maintain the right to curtail transmission capacity for maintenance. The severe penalties associated with not delivering contracted capacity, however, may prove a strong deterrent against signing reliability contracts.

An additional problem with allowing importing parties to sign reliability contracts is that contracts for interconnector capacity may have too short a duration. The reliability contracts may need to be signed several years in advance (depending on the design of the system) whereas most auctions of interconnector capacity only sell capacity for up to a year. Consequently, the same conclusion presents itself as before: the presence of significant, non-firm import flows either causes the solution to be expensive or its effectiveness is compromised.

Robustness against a regional shortage

In integrated systems, reliability contracts provide are robust with respect to regional shortages, if the system operator can limit purchases from outside the system.

Robustness against market power in the electricity market

The main improvement of reliability contracts over capacity requirements is that they provide operational incentives to generating companies for maximizing output during a shortage. For the volume of reliability contracts, a generating company's revenues are limited to the strike price, so there is no incentive to withhold power. On the other hand, the marginal revenues still are equal to the electricity price (as Example 6.3 demonstrated), so the higher the market price, the stronger the incentive is to increase sales.

Robustness against manipulation

A related advantage of capacity requirements is that by penalizing generating companies who sell reliability contracts that are not covered by operable generation capacity, capacity requirements provide an incentive to sell only contracts that are covered by available generation capacity, and by maximizing the available generation capacity during a period of scarcity.

An important vulnerability of reliability contracts is the possibility for generating companies to manipulate the contract auctions. In an equilibrium situation where the total volume of installed capacity is nearly optimal, the volume of reliability contracts to be auctioned would be about equal to installed capacity. This means that every generator with more than a very small market share will have market power in the capacity auction. A solution is to make sure that new market entrants are able to participate in the auctions (Vázquez et al, 2004). Therefore the success of this system is dependent upon the presence of low entry barriers in the generation market. It is questionable whether this condition is met in many electricity markets. (See the analysis in Section 5.6.) One prerequisite is that the auctions take place some years in advance, so newcomers have the time to construct new capacity if they are successful in the auction (Vázquez et al., 2002).

Problems with the exercise of market by withholding generation capacity occur in every market design: due to the long lead time for new capacity, withholding generation capacity may increase prices – be it in the electricity market, in an operating reserves market, in a capacity credit market or in an auction for reliability contracts. Similar solutions apply. As in the case of capacity requirements, it may also be possible to mitigate the generators' market power in a system with reliability contracts by using a price-elastic demand function for the reliability contracts. Allowing the use of interruptible contracts to cover some of the demand for reliability contracts would further contribute to the price-elasticity of the supply of the contracts.

Stimulation of demand price-elasticity

The incentive to consumers to exhibit their price-elasticity depends upon the specific details of the system. If consumers never need to pay more than the strike price, this clearly limits their price-elasticity in the same manner as a price cap. (If the strike price is high enough, this effect may not be very large.) The use of interruptible contracts as a substitute for reserve capacity may provide some additional incentives for demand-side management. Transaction costs are an obstacle to involving small consumers.

It may be possible to confront consumers with prices up to the average value of lost load, so demand price-elasticity would be as good as in an energy-only market. This would be achieved by returning the revenues, which the system operator obtains from calling the reliability options, not as a reduction of the electricity price during price spikes but spread out over time, for instance as a reduction of transmission tariffs. The net welfare effects would be the same (averaged over all consumers).

Supply-side efficiency

Reliability contracts have in common with capacity requirements that the total volume of

generation capacity is determined by a planning agent. The theoretical economic efficiency of a system of reliability contracts is the same as of a system of capacity requirements, as in both cases a planning agent determines the required reserve margin. In practice, the better operational incentives should enhance system reliability and avoid price manipulation during periods of scarcity.

Feasibility

As reliability contracts are a financial instrument, the implementation requirements are limited. A central planning agent needs to have the authority purchase the call options from the generators. A tracking system to allow trade of the options will also be necessary. In addition, a mechanism needs to be devised to pay for the options and to return the revenues, when the options are called, to the consumers. This system has not been tried in practice but detailed plans have been developed for implementation in Columbia.

Compatibility with a decentralized system

Reliability contracts were designed for an integrated system. In a decentralized system, the issue arises that a generating company that has sold a bilateral contract for its output no longer is hedged against the risk of selling an option contract. As its income is fixed, the risk of having to pay the market price minus the strike price when the option is called is uncovered. Section 8.2.2 discusses some possible solutions.

Another aspect of a decentralized system is that there may not be a way to limit sales outside the system. However, the options mean that all demand is hedged against high prices, which means that demand can always outbid any competing demand from outside the pool. This way, the options should guarantee the availability of an equivalent volume of generation capacity to the buyers of the options. This issue will be elaborated upon in Section 8.2.

Table 7-5: Evaluation of reliability contracts

Stabilization of investment	++
Effectiveness in securing generation in an open market	+
Robustness against a regional shortage	+
Robustness against market power in the electricity market	+
Robustness against manipulation	±
Stimulation of demand price-elasticity	+
Supply-side efficiency	+
Feasibility	+
Compatibility with a decentralized system	-

Conclusion

Reliability contracts provide an effective, market-oriented way to ensure generation adequacy. They combine many of the advantages of capacity requirements with better operational incentives to generating companies. However, again they may be vulnerable to capacity withholding, this time for the purpose of increasing the price of the reliability contracts. This may be limited by creating a price-elastic demand for the reliability contracts. Table 7-5 presents an overview of the advantages and disadvantages of reliability contracts.

The system of reliability contracts was designed for an integrated system. Some important changes to the design must be made to make it compatible with a decentralized system. The two main issues to be solved are how to achieve the targeted level of reliability in the presence of trade with energy-only markets and how to incorporate bilateral contracts in the capacity mechanism.

7.8 Capacity subscriptions

Stabilization of investment in an isolated system

Capacity subscriptions allow market forces to optimize the volume of available generation capacity. In this sense, they are a not only an effective capacity mechanism, but also promise to be the most efficient of the ones discussed so far. By directly linking available capacity to consumer peak demand (as estimated by the consumers themselves), the demand for capacity is made explicit, which simplifies making forecasts. Investment risk is further reduced by the steady revenue stream which the sale of capacity subscriptions generates.

A question is whether capacity subscriptions would not be vulnerable to the same cyclical effects that appear to threaten energy-only markets (see Chapter 5). During a period of excess generation capacity, competition would force the price of the fuses – the price for available generation capacity – to low levels, eliminating the incentive to invest. When demand starts to exceed available capacity, there is a time lag of several years before more generation capacity becomes available. There is a risk that in the intervening period the price of capacity increases to great heights, leading to over-investment, which would be followed by the next phase in the business cycle.

An important difference with energy-only markets is that the demand for capacity has the potential of being more price-elastic than the demand for electricity (especially without time-of-use metering). In a system of capacity subscriptions, a low price for fuses would lead consumers to cover much of their peak demand with fuses, as they would be able to purchase a high degree of reliability for a low price. As the excess generation capacity disappeared, less capacity would be available per consumer. A substantial portion of consumers would probably be willing to accept a smaller fuse, and risk having their consumption limited occasionally, if this would reduce their fixed costs significantly. Others would rather pay more than risk having to limit their consumption occasionally. Therefore the price of capacity would probably rise gradually as capacity became scarcer,

sending a more stable investment signal than that which arises from the price spikes in an energy-only market. Thus, demand for capacity probably would be fairly price-elastic – but only experience can tell to which extent this is true. The central question is whether the latent price-elasticity of demand is sufficient to dampen the tendency towards investment cycles.

A second question is how the consumer learning curve would develop. Starting in a position of excess capacity, consumers may not understand the value of having a certain amount of generation capacity reserved for them. They might not purchase fuses, or, if mandatory, minimally-sized ones. Thus, the generating companies would still not receive an efficient investment signal. Consequently, a shortage could develop, causing a price spike in the fuses, which could lead to an investment cycle. Consumers may not understand the system sufficiently to use it efficiently (Newbery, 2002b).

A final issue is that capacity subscriptions provide a firm limit to the consumption of electricity, whereas the supply of electricity has a stochastic nature because generator outages are partly unplanned. Therefore there is a possibility that total generation capacity is insufficient to meet the demand, even if the fuses are limited.[47] Additional outages will be caused by transmission and distribution network failures. Therefore the fuses do not shield consumers altogether from the risk of outages. This may be difficult to explain to consumers and hamper the acceptance of the system.

Effectiveness in securing generation resources in an open market

Capacity subscriptions offer a possibility of maintaining generation adequacy even in the presence of significant import flows. If the imports are not deemed reliable enough, the regulator may limit the sales of capacity subscriptions to domestic generators to ensure an adequate volume of generation capacity exists within the system.

Robustness against a regional shortage

Capacity subscriptions provide no guarantee that electricity is not exported during a regional shortage. The system ensures that supply and demand are matched physically but does not prevent the market from exporting the available electricity. Section 8.2.4 will consider possibilities for committing the output of generators who have sold capacity subscriptions.

Robustness against market power in the electricity market

Capacity withholding in the electricity market is unlikely, as generators have committed themselves to providing a certain amount of generation capacity. Due to the fact that demand is limited to available generation capacity, scarcity would be unlikely to develop, so prices much in excess of the marginal cost of generation would be suspicious. Only if unplanned generator outages cause the available generation capacity to be less than the

[47] It may be possible to reimburse consumers for part of their losses due to an outage through the penalty which should be levied upon generating companies who have less capacity available than they sold in fuses.

total capacity of all fuses – so service interruptions may still be necessary even after all fuses have been activated – would there be a reason for high electricity prices. To discourage capacity withholding to create or increase such price spikes, the penalty for not delivering as much capacity as has been sold through capacity subscriptions must be high enough. As the social cost of service interruptions equals the average value of lost load, this penalty should also equal the average value of lost load.

Robustness against manipulation

A system of capacity subscriptions offers no particular opportunities for the abuse of market power but generators with a large market share or an oligopoly of generators may still be able to manipulate the price of capacity. Withholding in the electricity market is possible, by selling fewer capacity subscriptions than the generating companies are able to. A high price for capacity subscriptions would attract new entrants to the market, unless the incumbents have excess generation capacity. However, if this is the case, there is a clear case for the competition authority that the generating companies are manipulating the price.

Stimulation of demand price-elasticity

Capacity subscriptions provide a strong incentive to all consumers to limit their peak consumption when capacity is scarce. Consumers receive an incentive to flatten their consumption pattern, as a result of which the relative height of system peaks will decline. This contributes directly to the overall efficiency of the electricity system, as less generation capacity needs to be available and existing capacity is used more efficiently.

Supply-side efficiency

Of the alternatives that have been reviewed until now, capacity subscriptions are the most economically efficient capacity mechanism. A significant source of economic inefficiency is removed by introducing a mechanism to allocate service interruptions based upon consumer preferences. However, there remains a need for government to establish a penalty for generating companies who cannot match their sales of capacity subscriptions with available generation capacity. The level of this penalty will determine the frequency with which generating companies are not able to meet their commitments, and therefore the reliability of the system. Thus, through the back door, government still has an influence upon the level of reliability.

There are two parameters that determine generation adequacy (given a certain demand): the total volume of available generation capacity that is desired, and the degree of certainty with which this volume of capacity actually is provided. In all other capacity mechanisms, the government establishes both; in a system with capacity subscriptions, consumers individually choose the volume, whereas government only needs to regulate the reliability with which this volume is obtained.

Feasibility

The main disadvantage of capacity subscriptions is that they require significant

adjustments to the system, mainly consisting of the installation of an electronic fuse at each consumer. However, the cost of such a fuse is less than that of a time-of-use meter. Consumers, especially smaller ones, need to be taught how to deal with this fuse: how much capacity they should purchase and how to reduce peak demand when the fuse is activated. In addition, the consumer acceptance issues that were discussed above (under 'Stabilization of investment in an isolated system') may present an obstacle to implementation. The novelty of capacity subscriptions requires extra attention to implementation. An option is to start with a pilot project among large electricity consumers, as they have most to gain from the cost savings which this system could provide, while the transaction costs would be relatively low.

Compatibility with a decentralized system

Capacity subscriptions are compatible with a decentralized system.

Conclusions

A system of capacity subscriptions is appears to be an effective way of ensuring generation adequacy. In theory, it is the most economically efficient of the proposed solutions because it allows them to select the reliability of service which they receive and confronts them with the cost of reliability. For generators, the advantage is that this system reveals the total volume of available capacity that consumers demand. A disadvantage is that this system requires a substantial investment in electronic fuses and a signaling system, and that it will take some time to implement. It may also not be effective in a system with strong connections to energy-only electricity markets, if there is no means to keep electricity from being exported. Table 7-6 summarizes the evaluation of capacity subscriptions. Finally, it is a question how consumers would respond to this system.

Table 7-6: Evaluation of capacity subscriptions

Stabilization of investment	+
Effectiveness in securing generation in an open market	+
Robustness against a regional shortage	--
Robustness against market power in the electricity market	+
Robustness against manipulation	+
Stimulation of demand price-elasticity	++
Supply-side efficiency	++
Feasibility	±
Compatibility with a decentralized system	+

7.9 Comparison

Overview

Table 7-7 combines Table 7-3 through Table 7-6 to present an overview of the performance of the different capacity mechanisms with respect to the selected criteria. The table also includes a row that shows whether the capacity mechanism has been tried in practice, as this is an important practical consideration.

Stabilization of investment

Table 7-7 shows that of the capacity mechanisms that have been tried in practice, only a system of capacity requirements is effective in dampening investment cycles. The two untried solutions, reliability contracts and the two versions of capacity subscriptions also promise to be effective. The reason is that these options directly control the volume of installed or available generation capacity, rather than influencing the investment equilibrium indirectly. The other options (capacity payments, a strategic reserve and operating reserves pricing) influence the volume of installed capacity only indirectly.

The first reason why price-based capacity mechanisms are less effective is what Oren (2000) refers to as the 'classic prices vs. quantities argument'. Because the demand curve for generation capacity has a steep slope and the supply curve has a gentle slope, a small error in the capacity price leads to a large shift in the equilibrium volume of generation capacity. Errors in controlling the quantity of generation capacity have a relatively small impact.

A related reason is that price-based mechanisms provide a less stable investment signal, as a result of which they are more susceptible to investment cycles, as Appendix A shows. Finally, the price-based capacity mechanisms require an estimate of the optimal volume of generation capacity as a basis for determining the optimal price level. Thus, an error in the estimate of the optimal volume of generation capacity is compounded by an error in the subsequent estimate of the optimal capacity price level. Precise knowledge of load-duration data and the average value of lost load are required to this end. As generating companies have difficulty estimating these data (in order to estimate price spike revenues), it is unlikely that a central planner would do better.

Effectiveness in securing generation resources in an open market

The presence of a substantial volume of imports reduces the equilibrium volume of installed capacity within the system. When imports are not firm, it may be decided to secure a sufficient volume of generation capacity within the system. Capacity payments may be directed to in-system generation capacity only; capacity requirements, reliability contracts and capacity subscriptions also may include requirements for either local generation capacity or imports with firm transmission rights. A strategic reserve and operating reserves pricing provide no means to control the location of new investments. To ensure self-reliance, the size of these reserves would need to exceed the volume of import capacity, which may mean an excessively large reserve.

Table 7-7: Comparison of the options

	capacity payments	strategic reserve, P_{sr}=VOLL	strategic reserve, P_{sr}<<VOLL	operating reserves pricing, P_{or}=VOLL	operating reserves pricing, P_{or}<<VOLL	capacity requirements	reliability contracts	capacity subscriptions
Stabilization of investment	-	--	±	--	±	++	++	+
Effectiveness in securing generation in an open market	+	-	-	-	-	+	+	+
Robustness against a regional shortage	-	-	-	-	-	±	+	-
Robustness against market power in the electricity market	±	--	±	--	±	±	+	+
Robustness against manipulation	±	+	+	+	+	±	±	+
Stimulation of demand price-elasticity	±	+	±	+	±	+	+	++
Supply-side efficiency	-	-	--	±	±	+	+	++
Feasibility	++	±	++	±	++	+	+	±
Compatibility with a decentralized system	+	+	+	+	+	±	-	+
Experience	yes	no	yes	no	yes	yes	no	no

Robustness against a regional shortage

An important problem with implementing a capacity mechanism in a strongly interconnected decentralized system, is that, despite the presence of a capacity mechanism, during a regional shortage the generators' output may be sold to neighboring systems. Absent a means to ensure that the electricity which is produced in the system is available to the consumers within the system, the latter would need to compete with consumers from outside the system. As a result, the electricity price and the reliability would be the same in all interconnected systems (barring network congestion). In this case, unilateral implementation of a capacity mechanism would be ineffective, even if it had secured sufficient generation capacity within the system to meet system demand.

Strongly interconnected electricity systems should implement a capacity mechanism jointly because it is more efficient and effective. If it is not possible to implement a capacity mechanism jointly in interconnected electricity systems, securing an adequate volume of generation capacity in individual systems is not sufficient. The availability of this generation capacity for the consumers who have paid for it is equally important. Only capacity requirements and reliability contracts provide possibilities to do this. However, these systems have been designed for a pool environment, where the pool operator may be able to control exports. An innovative solution is required for the decentralized systems of Europe. Section 8.2 discusses some options.

If the consequence of the current EU policy of leaving generation adequacy to subsidiarity (Art. 7, Directive 2003/54/EC) is that member states will implement different capacity mechanisms, this will likely distort international trade of electricity. It may also undermine the effectiveness and efficiency of the capacity mechanisms themselves. Moreover, it forces member states to experiment with innovative solutions. It may be concluded that generation adequacy should not be left to subsidiarity but requires regional coordination in order to be effective and to minimize inefficiencies.

Robustness against market power in the electricity market

One of the main issues that emerged from the analysis in Chapter 5 was the need to avoid the strong incentives for withholding generation capacity during shortages. None of the capacity mechanisms that have been tried in practice is fully successful in this respect. Operating reserves pricing and a strategic reserve reduce the incentive by lowering the maximum price, but the lower the maximum price, the larger the reserve needs to be. Capacity requirements do the same, if they are accompanied by a price cap. Only reliability contracts and capacity subscriptions provide positive incentives for maximizing output during periods of scarcity.

Robustness against manipulation

Care should be given that a capacity mechanism does not introduce new possibilities for manipulation, as was the case with capacity credits in PJM and New England (Hobbs et al., 2001b). The same solution applies to all systems that have some form of a market for generation capacity: administratively created demand for capacity should allow sufficient demand-elasticity to remove the incentive to withhold generation capacity. An

alternative, in the case of reliability contracts, is to auction the reliability contracts so far in advance that newcomers have time to enter the generation market.

The capacity mechanisms may also provide other opportunities for strategic manipulation. Price caps may be evaded by selling to an affiliate in a neighboring system and buying back, as was done in California during its crisis. Generation capacity that is not available may be sold under the assumption that it will not be called upon. During a regional shortage, available capacity may be sold to the highest bidder, which may be outside the system. Of each capacity mechanism, there are many variations possible. The final design should be carefully checked with respect to these and new forms of manipulation.

Stimulation of demand price-elasticity

The more effective capacity mechanisms (the last three in Table 7-7) have as an additional benefit that they also provide the best incentives for demand to become involved. Better demand price-elasticity would reduce the ratio between the peak and average volume of electricity consumption. This has the double advantage of bringing about a general efficiency improvement of the system and reducing the investment risk in peaking capacity.

Supply-side efficiency

All capacity mechanisms, except capacity subscriptions, require a central planner to determine the optimal volume of generation capacity. Given that this planner does not have perfect information (especially regarding the average value of lost load and the load-duration curve), this introduces a certain inefficiency. Energy-only markets, however, suffer partly from the same problem if demand price-elasticity is low. If there is a possibility that the supply and demand functions do not intersect, that is, if there is a possibility of service interruptions due to a lack of available generation capacity, there is a need for a maximum price. To determine this price, the average value of lost load must be known. Only capacity subscriptions allow consumers themselves to determine the optimal volume of installed capacity, so they are most efficient in theory. They are not perfect either, however, as a regulator needs to establish a penalty for generating companies who are not able to produce their contracted output.

Capacity payments that are accompanied by a price cap, a strategic reserve and operating reserves pricing all require knowledge of the stochastic distribution of the load-duration curve. In the absence of this information, they are likely to provide a wrong incentive. Thus there is a double error: first there is an error in establishing the optimal volume of generation capacity, then a second error is introduced because it is not known which combination of parameters (capacity payments or reserve requirements and price cap) leads to this volume. A strategic reserve finally introduces another source of inefficiency, which is that it disturbs the merit order of dispatch.

Feasibility

Unfortunately, the systems that are most easily implemented are the ones that perform

least well on the other criteria in Table 7-7. The administration and transaction costs of capacity requirements are substantial. Reliability contracts promise to be somewhat less demanding but the question is whether this is still the case when it has been adapted for a decentralized system. (See Section 8.2.2.) The cost of the fuses and consumer acceptance are the implementation barriers to capacity subscriptions.

Compatibility with a decentralized system

Most of the reviewed systems are compatible with decentralized systems. Unfortunately, the two most effective and feasible ones, capacity requirements and reliability contracts, have been designed for a pool environment. Capacity subscriptions are compatible with decentralized systems.

Other benefits of capacity mechanisms

When considering the implementation of a capacity mechanism, it should be taken into account that capacity mechanisms have other advantages than preventing possible failure of the generation market. Perhaps the most important advantage of having a volume of generation capacity that is theoretically optimal or even larger is that it reduces the opportunities for the exercise of market power. Stabilization of the business cycle has the additional advantage that prices become more stable and predictable. Capacity subscriptions have as an additional advantage that they resolve the issue as to how to determine the optimal volume of generation capacity. Finally, some of the capacity mechanisms provide incentives for developing the latent price-elasticity of demand, which would constitute an efficiency improvement.

Conclusion

The analysis shows the importance of aligning the economic and technical subsystems. For generation adequacy, financial incentives may not have the intended effect if they do not have a physical requirement attached. A strategic reserve or operating reserves pricing, for instance, may not lead to a higher reliability or more stable prices in an interconnected system. The next section will discuss this and other issues and how the most attractive capacity mechanisms can be adjusted to them.

7.10 Conclusions

This chapter provided a decision framework for selecting a capacity mechanism and described the advantages and disadvantages of several capacity mechanisms. In general, capacity mechanisms in which the volume of capacity is regulated or directly controlled by consumers are more effective at stabilizing investment cycles than systems that use economic incentives.

The most attractive capacity mechanism that has been tried in practice is PJM's system of capacity requirements. The fact that it was designed for an integrated electricity system should not present a significant obstacle to implementation in a decentralized system, if exchanges with other markets are limited. In an open, decentralized system it does not

appear robust against inter-system trade because price spikes from neighboring systems could be 'imported'. Its main disadvantage is that it provides insufficient incentives for generating companies to maximize the availability of generators.

The capacity mechanism called reliability contracts is designed to improve the operational incentives for generators but it is also tied to an integrated system. Possibilities for implementing it in a decentralized system will be explored in the next chapter. In the long-term, capacity subscriptions appear to be the most efficient solution but the implementation barriers are higher than for the other options and the practical effectiveness is unproven. An option is to use this system only for large consumers, whose service can be interrupted individually (so the reliability of their electricity service can be controlled independently from the rest of the system).

None of the proposed capacity mechanisms appears robust in decentralized systems with significant out-of-system trade, such as most European markets. Therefore the current EU policy of leaving generation adequacy to subsidiarity should be replaced with stronger regional coordination; preferably, the same capacity mechanism should be implemented simultaneously in as large a group of interconnected electricity systems as possible. If this is not feasible, individual member states who wish to take measures to safeguard the volume of generation adequacy and its availability during regional shortages will need to develop innovative solutions. The next chapter discusses some possibilities.

8 Generation adequacy in Europe

The analysis in the previous chapter concluded that there is no satisfactory capacity mechanism that can be implemented in a decentralized, interconnected electricity system such as most European markets, unless it is introduced in all connected markets. This chapter explores options for adjusting the existing proposals for interconnected, decentralized systems. In addition, this chapter provides an overview of the main policy choices regarding generation adequacy.

8.1 Introduction

It was shown in Chapter 7 that implementation of a capacity mechanism is easiest in a relatively isolated system, such as the U.K. and Ireland, the Iberian peninsula or Greece.[48] For more strongly interconnected systems, the ideal clearly is the joint implementation of a capacity mechanism by all interconnected systems, so the mechanism can be effective without distorting inter-system trade. The question is whether it is politically feasible to implement a capacity mechanism jointly in many electricity systems.

Strongly interconnected systems may not have the time to wait for the regional implementation of a capacity mechanism. If they see an investment cycle looming, they may desire to secure their generation adequacy independently. While this is a second-best option, because it will reduce the economic efficiency of the interconnected system, it may still be preferred to risking a reduction of the reliability of service. It may be considered a temporary safeguard measure while a regional solution is being developed.

[48] Few regions are entirely closed, but if the interconnection of the entire region is small, relative to electricity consumption within the region, the effect of exchanges upon the effectiveness of the capacity mechanism may be small. This may be the case for Nordel, the UK or the UCTE as a whole. In the latter case, however, the large size of the interconnected region makes it clear that such a significant change to the market structure as the implementation of a capacity mechanism will not quickly be agreed upon, especially considering the already substantial differences between the member systems. In the UCTE, individual systems may consider it necessary to implement a capacity mechanism on their own before regional consensus has been reached.

In this chapter the possibilities are analyzed to implement variants of the capacity mechanisms that were found to be most attractive in Chapter 7 (capacity requirements, reliability contracts and capacity subscriptions) to the specific conditions of interconnected decentralized systems such as most European electricity markets.

Section 8.2 will start by exploring options for adjusting the known capacity mechanisms for open, decentralized systems. Section 8.3 describes the policy choices that are to be made with respect to generation adequacy and summarizes them in a decision tree. Section 8.4 discusses some implementation issues. Sections 8.5 and 8.6 present the conclusions and recommendations for European markets.

8.2 Innovative capacity mechanisms

8.2.1 Introduction

The analysis in Chapter 7 showed that the most attractive options in the short term are capacity requirements and reliability contracts. Both capacity mechanisms were designed for an integrated electricity system. Implementation in a decentralized system may require some adjustments, however. Implementing capacity requirements in a closed, decentralized system should not cause serious complications, compared to the variant for an integrated system, as the mandatory pool is not crucial to the effectiveness of this capacity mechanism. In an open decentralized system, however, it may not be possible to 'recall' exports from generating companies who have sold capacity credits. It may be possible to require them to sell to a power exchange or balancing market but decentralized systems do not appear to provide the electricity from being (re)sold to buyers in neighboring systems. To ensure that the consumers within the system, who are the ones who pay for the capacity credits, also are the ones to benefit from any enhancement of reliability, the capacity mechanism needs to be adjusted. When supply is tight, the electricity that is produced by generators who have received capacity payments must be made available to the consumers within the system. Sections 8.2.2 and 8.2.3 discuss some possible solutions.

If real-time meters are present, this presents an opportunity for an innovative version of capacity subscriptions, which will be discussed in Section 8.2.4. Here, too, an important issue is how to accommodate inter-system trade.

8.2.2 Reliability contracts in an open, decentralized system

Two options for implementing capacity requirements in open, decentralized systems present themselves, which will be called the physical and the financial variants. In both variants, generating companies are allowed to sell both call options and contracts in the bilateral market or in the power exchange.

Physical variant

In the physical variant, a generating company that sells an option commits itself to

offering all capacity that underlies the option and that is not committed through bilateral contracts to the balancing market at or below a given strike price.[49] So if a generating company has sold 500 MW worth of options and has contracts to produce 400 MW, it is required to offer the remaining 100 MW to the balancing market at the strike price. The system operator simply needs to compare each generating company's schedule for the next day to its option volume to know how much capacity he can call. If the generating company is not able to produce 500 MW, it needs to pay the difference between the market price and the strike price to the system operator for the volume of capacity that it cannot produce, the same as in the original reliability contracts proposal.

The question is whether this proposal can be made robust in a decentralized electricity system with strong trade with energy-only markets. The system operator does not see how much generation capacity is committed to exports, other than by observing the daily export schedules. He does not know the future export obligations of generating companies, and can therefore also not determine whether the available generation capacity will be sufficient to meet demand. There is no mechanism to ensure that sufficient generation capacity will be reserved for consumers in the system. Therefore this option, in its current form, appears less attractive for an open market. In a system with relatively little interconnector capacity, its simplicity makes it an attractive option.

Financial variant

In the financial variant, generating companies who have sold reliability contracts are required to pay the market price minus the strike price for the entire volume of options, if they are called (Vázquez et al., 2004). To the extent that the generating company's output was committed through bilateral contracts, the option payments would constitute a significant loss. The generating company would be compensated for this loss through parallel contracts that return the option payments to the degree that the company was producing for bilateral contracts.

Consider, for example, a generating company that has a base-load contract for 400 MW at a price of 25 €/MWh and no other contracts. The revenues from the base-load contract are 10,000 €/h. Assume that the company has sold call options for 500 MW and the strike price is 1000 €/MWh. Assume a demand peak develops during which the spot price rises to 1200 €/MWh. The system operator calls the options, which means that the generating company is required to pay the difference between the spot price and the strike price times the volume of options that he sold, which is 200·500=100,000 €/h. The company's revenues are 10,000 €/h plus the revenues from selling 100 MW in the spot market, equal to 120,000 €/h. Because 400 MW of the company's output does not receive spot prices, the system operator returns the corresponding option payments: 200·400=80,000 €/h. As a result, the net option payments from the company to the system operator equal 20,000 €/h, equivalent to the option payments for the 100 MW that was sold on the spot market. Consequently, the company's net revenues are equal to 25 €/MWh for the part of its output that was sold through a bilateral contract and 1000 €/MWh for the part sold in the spot market.

[49] This is the author's interpretation of a proposal by Henney and Bidwell (2003).

Effectively, this system limits generating companies' requirements to make option payments to the volume of generation capacity that is not contracted under long-term contracts. The volume of option contracts still is equal to the total volume of generation capacity but the incentive effect of the options only applies to the part of generation capacity that has not been committed through long-term contracts. In a closed system, the effect is the same as for the physical variant: the generating company receives the strike price for all capacity that was not sold through a bilateral contract and pays the market price minus the strike price for unavailable capacity, when the options are called. A drawback is the added complexity of parallel contracts and the large associated financial flows.

An advantage is that this capacity mechanism appears more robust against a regional shortage. While there still is no way to ascertain that bilateral contracts are not sold outside the system, the system operator can call options for the entire volume of demand. This means that the load-serving entities can bid any price, if necessary, to obtain enough electricity for their consumers, as they are fully hedged against price spikes. This should allow them to out-bid competitors from neighboring energy-only markets.

8.2.3 Bilateral reliability contracts

Another possibility is to require the load serving entities, rather than the system operator, to purchase the reliability contracts. This would effectively entitle them to electricity at a certain price.[50] This capacity mechanism will be dubbed 'bilateral reliability contracts'. It is based upon the same principle as the original, 'central' reliability contracts proposal: load-serving entities are required to purchase call options from generators for a volume that is determined by the regulator, based upon their peak consumption plus a reserve margin. The requirement is backed by a penalty for load-serving entities who have not purchased a sufficient volume of reliability contracts. The reliability contracts are registered by a central agent, such as the system operator or the regulator, to ensure that the generating companies only sell a volume of contracts that they can back with generation capacity and that the load-serving entities purchase a sufficient volume of reliability contracts to meet their obligation. To verify that parties keep to their contractual obligations, use can be made of the already existing systems in which the system operator is notified of scheduled generation and consumption (Knops, 2003).

This proposal resembles the system of capacity requirements in PJM in as much that a bilateral capacity market develops. The crucial difference is that when a generating company sells bilateral reliability contracts, it commits to offering its output to the load-serving entity that has purchased the option, rather than to a power pool. This means that the load-serving entities who pay for the generation capacity by purchasing reliability contracts have access to the generation capacity when they need it at the strike price. In an open, decentralized market, this solves the problem of 'leakage' of capacity to neighboring systems. The load-serving entity may choose to purchase electricity

[50] A version of this option was introduced by Oren (2000) as an improvement upon capacity payments. Both Shuttleworth et al. (2002) and Vázquez et al. (2004) mention this option briefly. Hobbs et al. (2001c) also propose amending PJM's system in this manner.

elsewhere if the market prices are low but retains the choice to call the option when that is more attractive.

Vázquez et al. (2003) also consider a bilateral variant of reliability contracts but consider it less attractive – at least for the Netherlands – than the centralized variant. However, their main objection is that the revenues of the Dutch load-serving entities are regulated, which is only the case until July 1st of 2004, after which time there will be full retail competition. By July, 2007, all EU member states will be required to allow full retail competition (Art. 21, Directive 2003/54/EC). In the remainder of this analysis, it will be assumed that the market for which this capacity mechanism is developed is an open, decentralized electricity system with full wholesale and retail competition.

As in the case of capacity requirements and reliability contracts, the option premium can be left to the market. To a degree, it may also be left to market parties to decide the strike price. However, an option with a strike price equal to the average value of lost load provides no risk mitigation and therefore has a value of zero. This would render the requirement to cover expected demand plus a reserve margin with option contracts moot. Therefore the regulator would need to set a maximum strike price. Generators and load-serving entities would be free to choose lower strike prices; a strike price of zero would turn the option contract into a regular energy contract.

Similarly, the regulator would need to choose a minimum contract duration to prevent the option contracts from converging with spot contracts. The experience in PJM teaches us that the minimum contract duration should be at least a number of months so generators cannot change their positions during a price spike (Hobbs et al., 2001c). The length of the contracts does not need to impede short-term efficiency. If a load-serving entity holds a volume of options that are in the money in excess of its own demand for electricity, it can resell the surplus. If a generating company's options are not called while the market price exceeds the company's variable costs (which means that the options have a strike price higher than the company's variable costs), the company may produce for the spot market. Thus efficient short-term allocation is achieved. In case of a shortage, a load-serving entity's risks are limited, as the call options that it owns guarantee that it has enough capacity available at the strike prices of the different option contracts. See Example 8.1.

An important advantage of bilateral reliability contracts is that dependence upon the government to design an efficient auction is replaced by reliance upon a number of competing market parties (load-serving entities) to purchase the option contracts. The latter probably have better knowledge of the market than the government does and they have a strong incentive to do all they can do to reduce the price of the contracts. To attract newcomers, they may also sign multi-year contracts or purchase reliability contracts a number of years in advance. This solves the question in the centralized variant of reliability contracts of the timing and duration of the auctions: the load-serving entities themselves can find the optimum between liquidity and contract duration. In addition, if the load-serving entities cannot find reliability contracts at an affordable rate, they may invite their customers to sign interruptible contracts or develop generation capacity themselves. This further limits market power in the capacity market.

A second advantage is that the financial flows are significantly simpler than in the financial variant (for decentralized markets) of the centralized reliability contracts scheme that was discussed in Section 8.2.2. Rather than having parallel bilateral contracts and reliability contracts, with the need to reimburse holders of long-term bilateral contracts when their options have been called, bilateral reliability contracts may substitute for the long-term bilateral contracts. If the strike price of an reliability contract is set to zero, the option premium becomes the only payment and the reliability contract is equivalent to a regular bilateral contract. The only limitation is that this contract would need to have a minimum duration in order to qualify as a reliability contract. To the extent that the load-serving entities do not want to commit themselves in long-term contracts, they could purchase reliability contracts with a strike price equal to the maximum, which would have a lower premium. Then they would purchase electricity in short-term markets and only call the option during price spikes.

Example 8.1: Bilateral reliability contracts

A load-serving entity has customers with an expected annual peak demand of 4000 MW. The regulator requires it to purchase reliability contracts for a capacity equal to its peak demand plus a reserve margin. If it is assumed that the reserve margin is set at 10%, the load-serving entity is required to purchase reliability contracts for a total of 4400 MW. The load-serving entity itself owns 2000 MW of generation capacity. If this capacity is rated at 90% availability, the load-serving entity needs to purchase 4400-1800=2600 MW worth of reliability contracts. To the extent that the strike price of the options is below the current market price, the load-serving entity will call them. The load-serving entity will purchase the rest of its demand on the market.

Bilateral reliability contracts are compatible with inter-system trade because they contain a requirement to deliver to a specific load-serving entity. Therefore the incentive is removed for generating companies who have sold reliability contracts to export at prices above the strike price during a regional shortage. It would still be allowed but consumers could buy the electricity back, as they are hedged against prices in excess of the strike price.

As in the New York system of capacity requirements (Hobbs et al., 2001c), a requirement can be included that certain volumes of reliability contracts be procured from generators in specific parts of the network. This would allow the reliance upon imports to be controlled and it could also be used to ensure a balanced geographical development of the generation stock.

The similarity to PJM's capacity requirements means that some of the same difficulties may be expected. One issue is how to estimate each load-serving entity's share of the total capacity obligation. PJM appears to have found a satisfactory solution in its estimation methods (PJM, 2001). An option requirement does not change the issue of market power in the capacity market, such as PJM has experienced. The administratively determined demand for option contracts (for capacity credits in the case of PJM) leads to highly volatile prices and provides an easy venue for the exercise of market power. Stoft

(2002) suggests to make demand more price-elastic. (The penalty should increase with the deviation from the required volume of options.) On the other hand, options should solve the 'iron in the ground' problem experienced in PJM, the problem that generation capacity may not be available in real time, in the same way that reliability contracts do.

A question is how to implement bilateral reliability contracts in a market in which generating companies are vertically integrated with retail companies. Presumably, the company's generation capacity would be subtracted from its obligation to purchase reliability contracts. If the holder of the options would be the retail section of the company that sells the options, any penalties for not being available would remain within the company. Consequently, the 'iron in the ground' problem resurfaces: the firm would have an incentive to overstate the availability of its generation assets in order to reduce its obligation to purchase reliability contracts.

It may therefore be necessary to rate the availability of the company's generating assets, like PJM does. However, care must be given not to reduce the incentive to generating companies to maximize their output during shortages, as this is one of the main advantages of reliability contracts. A solution may be to allow the generation companies to rate the availability of their own generators and to declare this to the system operator, who issues a penalty when the availability during shortages is less than stated. If the penalty is high enough, for instance equal to the spot price of electricity minus the maximum option strike price, the incentive to maximize output would be similar to the incentive that an independent generator would receive from selling reliability contracts.

Another significant issue is the question of counterparty risk: the risk that generating companies may not be able to make the option payments if the availability of their generators is less than their option volume. This risk could be reduced by limiting the sale of reliability contracts to each generating company's installed capacity.

Conclusion

Bilateral reliability contracts should be robust against a regional shortage, are compatible with a decentralized system and do not require auctions. They are a straightforward way of achieving reliability because they create direct contractual connections between the load-serving entities and the generating companies for the entire demand plus reserve margin. The main downside is that their effectiveness may be limited by vertical integration of generators with retail companies.

8.2.4 A financial version of capacity subscriptions

Implementation in a closed system

In some electricity markets, commercial customers already have time-of-use electricity meters. These meters provide a possibility of creating a financial version of capacity subscriptions. Rather than limiting peak demand with a physical instrument such as an electronic fuse, the author suggests the use of financial instruments that are based upon the time-related metering data. This would lower the implementation costs, compared to the original capacity subscriptions, as the electronic fuses would not be necessary. In

addition to time-of-use meters, there needs to be a signal to let consumers know when they need to limit their consumption to their contractually agreed maximum. Two questions arise: what type of contract could replace a capacity subscription while providing the same efficient incentives to generators and consumers? Second, would parties in a liberalized market engage in these contracts voluntarily or would there need to be a need for regulation?

A simple type of electricity contract with a capacity incentive consists of a capacity component, which is related to the consumer's highest peak in consumption, and an energy component that reflects total electricity consumption. This type of contract offers only a weak incentive to consumers to manage their peaks during periods of scarcity. If the capacity component is a fixed or even declining price per unit of capacity, the private cost to the customer of increasing his peak demand is less than the social cost if the system is short of capacity. As the incentive for consumers to limit their consumption to a certain level is insufficient, generators also receive a insufficient signal regarding the need for peaking capacity. In addition, a shortage of generation capacity results in high electricity prices, so generators are collectively rewarded rather than punished for a capacity shortage. Thus, it appears that contracts with a simple capacity component provide neither generators nor consumers with sufficient incentives solve the issue of generation adequacy.

A financial version of capacity subscriptions would entail a requirement for consumers to limit their consumption to a pre-stated amount, similar to the original proposal. They would be free to choose this limit but they would have to pay for it and commit to staying beneath it when the capacity reserves in the system are low. In the original version of capacity subscriptions, a physical device is used to limit consumption when necessary. The financial version would consist of a contract with a penalty for exceeding the limit. As in the previously discussed cases, the penalty needs to be elastic to mitigate market power in the capacity subscription market.

Figure 8.1 shows the structures of the different types of contracts. The vertical axis shows the price per unit of capacity, while on the horizontal axis is electricity consumption. In the original system of capacity subscriptions, the system operator can apply a physical limit to each consumer's use of electricity. In the figure, this is indicated by the fact that the price curve for capacity ends in a vertical line (indicated with '*a*'). A consumer who has a fuse with a size of q_c *kW* pays a price of $P \cdot q_c$.

The financial variation, indicated by line *b*, has nearly the same result: consumers can exceed their chosen peak capacity but at a penalty, which increases with the degree to which the contracted volume is exceeded. In this case, the consumer pays $P \cdot q_c$ plus a penalty which is determined by the penalty function $f_{pen}(q-q_c)$. In both cases, consumers would have the option of increasing their limit by purchasing more capacity in advance.

These systems contrast with regular contracts, in which the marginal cost of peaking capacity stays constant, as is indicated by line *c*. In this contract, consumers pay a fixed component equal to their peak consumption, $P \cdot q_c$ but they keep paying the same price P for a higher peak consumption. This type of contract provides generators with much less

certainty regarding the demand peak they can expect.

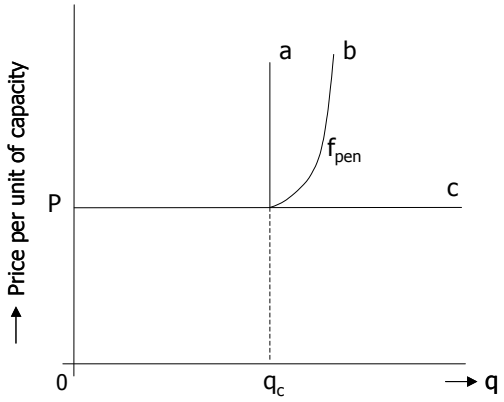


Figure 8.1: The price of capacity in different types of contracts: a) physical capacity subscriptions; b) financial capacity subscriptions; c) contract with fixed capacity payment

To avoid free-riding in the manner described in Section 5.5.2, the financial capacity subscriptions would need to be mandatory. An independent agent needs to monitor and enforce the financial capacity subscriptions. This agent would register the volume of capacity subscriptions sold by generators and to whom they are sold. The agent would also have the authority to penalize consumers who exceed their contracted peak consumption and generators who do not deliver the capacity that they have sold, at least during periods when the electricity system is short of capacity. To eliminate the negative externality of random service interruptions, the penalties would need to equal the cost of the random service interruptions caused by generators over-selling capacity or consumers exceeding their contracted peaking capacity. Thus, the penalty should rise to the average value of lost load during shortages.

A financial version of capacity subscriptions should have an impact similar to the original proposal with electronic fuses. The main negative difference is that its effectiveness depends upon the enforceability of the contracts. The main advantage is that this system is easy to implement in a system in which consumers already have time-of-use meters, as there would not be a need to install an electronic fuse at each consumer.

Table 8-1: Overview of the options in an open, decentralized system

	Reliability contracts, physical variant	**Reliability contracts, financial variant**	**Bilateral reliability contracts**	**Financial capacity subscriptions**
options purchased by:	central agent (TSO?)	central agent (TSO?)	load-serving entities	end consumers
option volume:	system peak demand plus reserve margin			determined by consumers
option price determined by:	central agent (TSO?)	central agent (TSO?)	market players	consumers
status of physical bilateral contracts:	allowed	allowed	the option contracts are the only physical bilateral contracts (which need to have a minimum duration to qualify)	the option contracts are the only physical bilateral contracts
treatment of physical bilateral contracts:	production for a physical bilateral contract satisfies option requirements when called	option payments are returned to the extent that output was covered by bilateral contracts	the options are called by the load-serving entities	the options are called by the consumers

(Table continues on the next page.)

(Table continued from the previous page)

	Reliability contracts, physical variant	**Reliability contracts, financial variant**	**Bilateral reliability contracts**	**Financial capacity subscriptions**
financial flows:	• from consumers via the central agent (to finance the options) to generating companies (to purchase the options) • from generating companies via the central agent (when options are called) back to consumers (revenues from calling the options) • from consumers via load-serving entities to generating companies	the same as in the physical variant, plus a restitution from the central agent to the generating companies for the option payments that were made for generation capacity that was committed in bilateral contracts	only from consumers to load-serving entities and from them to generating companies	only from consumers to load-serving entities and from them to generating companies
robustness in a system connected to energy-only markets:	not robust: the bilateral contracts may be sold outside the system	robust: the options protect consumers against price spikes, so the load-serving entities can out-bid buyers from outside the system on behalf of their consumers	robust: the load-serving entities secure a volume of capacity and a price limit for themselves	robust: the load-serving entities secure a volume of capacity and a price limit for themselves

Implementation in an open system

The above proposal still suffers from the main disadvantage of physical capacity subscriptions, which is that they are not robust against a regional shortage. Capacity and energy sales are not linked, so generation adequacy within the system does not mean that the generation is also available to the purchasers of the capacity subscriptions. Again the solution appears to be to turn the capacity subscriptions into option contracts. The difference with the previous proposals is that the end consumers would purchase the options, and would also commit to not consuming more than the option volume during a shortage. This way, the generating companies would commit to selling electricity to the buyers of the capacity subscriptions, which should make the system robust to a regional scarcity. In a system with real-time meters, this appears to be the most elegant way to secure generation adequacy, as it allows consumers the choice of how much capacity they need.

Again, the conclusion is arrived upon that a kind of reliability contract is needed to secure generation capacity in an open, decentralized system. The difference with the reliability contracts described in the previous section is that in the proposal at hand individual consumers would be the holders of the options. The presence of real-time meters makes this sophisticated type of consumer contract possible.

8.2.5 Overview

Table 8-1 provides an overview of the innovations to the existing capacity mechanisms that were discussed in this section.

8.3 Policy choices

The previous sections discussed a number of capacity mechanisms. Which one should be implemented, in particular in a decentralized system? Should a capacity mechanism be implemented now or should we wait until we have more evidence of the dynamic nature of electricity markets? This section discusses the policy choices, with a focus upon European electricity systems.

8.3.1 Implementation as a precaution?

The first question to be resolved is whether there really a need to implement a capacity mechanism, or should we wait and see how the market develops, considering the lack of empirical evidence of market failure?

> *Policy choice 8.1: Should a capacity mechanism be implemented preventively, which is easier but for which the need has not been demonstrated, or should it only be implemented when the need is clear, which means that reliability may be jeopardized for some time and the transition phase may be more difficult?*

Waiting entails a significant risk, as it is not possible to monitor the market and forecast

generation adequacy with sufficient certainty, far enough into the future, to allow time for intervention when it becomes apparent that a shortage of generation capacity looms. Not only does the development and implementation of the capacity mechanism take time, the industry will also need time to evaluate its implications in order to adjust investment strategies and, last but not least, it will take time to construct additional generation capacity in response to the new capacity mechanism.

If, in the mean time, the volume of generation capacity drops below the level that the capacity mechanism is designed to obtain, a difficult transition period will follow. The transition period is apparent in the model in the Appendix, especially in the runs with a higher growth rate of demand. During this period the reliability of service will be lower than desired, while the inability of the market to immediately provide the desired volume of generation capacity may cause high capacity prices in a capacity market or in a system with capacity subscriptions.

Implementation of a capacity mechanism during a period of excess capacity, on the other hand, is much easier, as it would not require an immediate physical reaction from the market. The market could continue to reduce the capacity margin until the limits of the capacity mechanism would be reached, after which it would stabilize. The smoother transition and the lower risk to the reliability of service are arguments in favor of such a 'preventive' strategy.

Politically, however, the balance may shift in the other direction. Implementation of a capacity mechanism is a significant intervention in the electricity market. Without a clear sense of urgency, it may be difficult to gain support, both political and from the sector, for such a change. The expected social costs of not taking any action likely are much higher than the implementation costs of a capacity mechanism. However, due to the long time scale at which the generation sector develops, the resulting political repercussions will probably not affect the political leaders who currently are in office.

8.3.2 Unilateral or regional implementation?

Strongly interconnected electricity systems face the question as to whether to implement a capacity mechanism themselves, or whether to try to find a regional solution. The latter is not only the more efficient solution, it also is easier and there are more suitable capacity mechanisms available. A system of capacity requirements, for instance, can be implemented without much difficulty in a decentralized system if there are no significant imports and exports. If there are, an adapted version needs to be chosen, such as the financial variant of reliability contracts that was described in Section 8.2.2 or the bilateral variant that was proposed in Section 8.2.3. However, the regional development of a capacity mechanism may take much time, especially in a network with as many constituting systems as the UCTE. There may not be enough time to develop a regional solution before the first investment cycle develops. A dilemma is the consequence:

> *Policy choice 8.2: Should importing systems implement a capacity mechanism unilaterally, despite the distortion of the greater market, or rely on imports and hope that a collective solution will be developed in time?*

In Europe, the Netherlands and Italy import a large part of their electricity. Therefore these countries face the difficult choice between unilateral implementation of a capacity mechanism, which would be more expensive or less effective, or waiting for a European solution, which may take too long.

An additional disadvantage of unilateral implementation is that if, eventually, a regional solution is devised, the individual capacity mechanisms would need to be replaced. Long-term commitments that generating companies and/or load-serving entities had engaged in under the first capacity mechanism could become stranded investments in the new capacity mechanism. This is an argument for systems in which the system operator is the only buyer, such as in a strategic reserve, operating reserves pricing or the central version of reliability contracts.

8.3.3 Self-reliance?

If the choice for solitary implementation is made in a system with strong interconnections, another question immediately presents itself. Should physical self-reliance be the goal? Alternatively, to which degree can imports be relied upon in the long term? The issue is not only the physical availability of imports but also the price at which they are available. Capacity mechanisms tend to reduce the price volatility of electricity markets; some provide an upper limit to the payments for energy. Imports from energy-only markets could undo this effect as they would cause price spikes in neighboring systems also to be imported, which would leave consumers to pay both for the capacity mechanism and for price spikes. This would undermine one of the main advantages of having a capacity mechanism.

For these reasons, it may be chosen to become self-reliant, if neighboring systems do not implement a similar capacity mechanism. The cost of self reliance may also be high, however, for electricity systems with a large share of imports. They may compromise by requiring a lower reserve margin in their capacity mechanism than would be considered optimal in an isolated system. While this reduces their security of supply to the extent that the imports are not dependable, it also reduces the cost of supplementing these imports with presumably inactive back-up generation.

> *Policy choice 8.3: If an interconnected electricity system chooses unilateral implementation of a capacity mechanism, should it become fully self-reliant? If not, to what degree should it depend upon imports?*

8.3.4 Innovativeness

The evaluation in Section 8.2 showed that the more innovative variants of reliability contracts promise to be more effective, in particular in open, decentralized systems, but the lack of experience casts some uncertainty upon their practical merits. Theoretically, they should provide better incentives to generating companies and be robust with respect to inter-system trade. However, the vulnerability of these untried systems to gaming, for instance, is unknown.

> *Policy choice 8.4: Should a capacity mechanism be chosen that has been tried in practice but has known flaws, or should the choice be made for a more innovative system with better theoretical incentives but unknown flaws?*

In the case of unilateral implementation in a decentralized system with strong interconnections, the only choice is to implement the one of the innovative variants of reliability contracts. The alternatives are to do nothing (and perhaps trying to achieve a regional solution) or to implement a capacity mechanism of which the effectiveness is uncertain. In an integrated system, PJM's system of capacity requirements may be implemented, also if it has strong interconnections.

8.3.5 Short-term versus long-term options

The choice of capacity mechanism depends upon the specific circumstances of the system within which it is to function. If a capacity shortage already is looming, it may be necessary to implement a capacity mechanism that can be implemented quickly, as a transition measure, even if it does not meet all the criteria. Capacity payments, a strategic reserve and operating reserves pricing are relatively easy to implement, which makes them attractive as short-term solutions. Unfortunately, their effectiveness is limited and they entail a risk of distorting investment incentives. Whether to implement a short-term solution is a judgment call: if it is estimated that enough time remains to develop a more elaborate but also more effective and efficient capacity mechanism, this will be preferable.

> *Policy choice 8.5: Should a capacity mechanism be chosen that can be implemented quickly, or one that requires more implementation time but probably also more effective and efficient?*

If the decommissioning of old units threatens the capacity margin, the system operator may choose to purchase them as a strategic reserve (providing he has the authority to do so). Creating a strategic reserve this way was Sweden's response to concerns about generation adequacy in recent years. An alternative that can be implemented just as easily is operating reserves pricing. Expanding the operating reserves when the reserve margin is below the target level would immediately create an investment signal. The disadvantages of these methods are that their effectiveness in stimulating investment is uncertain, that they mitigate but do not eliminate the problem of capacity withholding in the electricity market, and that they are not robust against regional shortages. Therefore they should only be considered as temporary solutions. If more time is available (on the order of five to ten years before a shortage is projected), either capacity requirements (in a closed decentralized system or an integrated system) or one of the options that were discussed in Section 8.2 are more effective and efficient.

8.3.6 Overview of the policy choices

The above policy choices are summarized in Figure 8.2. The diagram shows the consecutive choices that present themselves as well as to which capacity mechanisms they lead. The first choice to be made is whether a capacity mechanism will be

implemented right away, as a precaution, or only when it becomes clear that the market is not providing sufficient generation capacity. Chapter 5 concluded that waiting is a risky policy because failure of the market to provide sufficient generation capacity cannot be predicted far enough in advance to allow time to implement a capacity mechanism.

The next issue is whether regional implementation is feasible, as this is preferable to implementation by individual systems within a larger interconnected network. If regional implementation of a capacity mechanism is not likely to happen in time to avoid a shortage, individual systems may decide to take action. If they are weakly interconnected, the absence of regional measures does not matter much, as all options are still open.

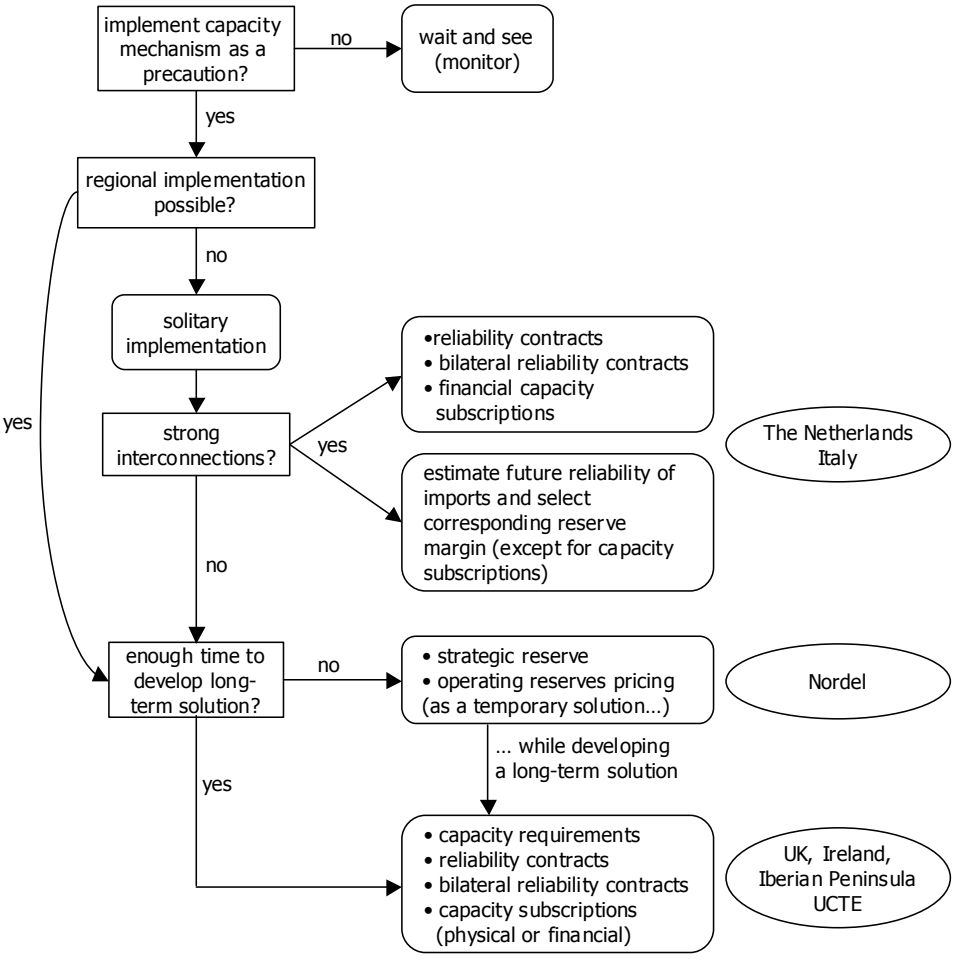


Figure 8.2: Decision framework for implementation of a capacity mechanism in a decentralized system

Unilateral implementation of a capacity mechanism by a strongly interconnected electricity system is a difficult issue. None of the capacity mechanisms presented in Chapter 6 appears robust to a regional shortage, except the two pool-based systems (capacity requirements and reliability contracts). This means that in a decentralized system one of the innovative capacity mechanisms of Section 8.2 needs to be implemented.

The ovals on the right hand side of Figure 8.2 indicate which options are available to which European countries. The UCTE as a whole is so large, relative to the exchanges with its neighbors, that it can be considered as an isolated system. Examples of other relatively isolated systems are the UK, Ireland and the Iberian Peninsula. These markets are in the comfortable position of having all options available because they are not under pressure to develop a temporary solution, nor do they have such strong interconnections that they are confined to the options for open markets.

Development of a capacity mechanism at the level of the UCTE, or even for part of it, may take too long for some member systems. The Netherlands and Italy are the largest importing countries in the EU (UCTE, 2002c), which may given them cause for concern with respect to future reliability. If these countries would decide to implement a capacity mechanism independently, they would need to choose one that is robust with respect to imports and exports. In a decentralized system, an attractive option appears to be a system of bilateral reliability contracts (Section 8.2.2).

In Nordel, the shortages in Sweden and Norway have prompted implementation of a strategic reserve and a form of operating reserves pricing, respectively, as short-term solutions while a longer-term solution is being developed.

8.4 Implementation issues

Adaptation to local conditions

If the choice is made for capacity requirements, the ample experience in the PJM system provides the opportunity for empirical study of this capacity mechanism. Much has already been written about PJM's ICAPS, however, when considering implementation of capacity requirements in another system, the potential impacts of the differences between the two systems should be assessed. For instance, large imports have not been an issue in PJM.

Unilateral implementation of an effective and efficient capacity mechanism in one of Europe's decentralized systems requires innovation. None of the available systems are fully satisfactory. The most promising option that appears feasible in the current institutional and technical setting of most markets is a system of reliability contracts (central or bilateral).

New systems must be thoroughly tested with respect to their ability to stabilize the generation volume in the presence of insufficient information regarding future supply and demand conditions and risk-averse behavior by both producers and consumers. In

addition, they should be robust against market power among generating companies, both in the short and the long term. Thus, a combination of system dynamics and market power modeling is required. Ford (1996) used a system dynamics model without market power. A similar approach was used in the Appendix, where a first assessment was made of the dynamic behavior of several capacity mechanisms. Future research should focus on the inclusion of strategic behavior in the models. With respect to generation adequacy, the long-term dynamics of less than perfectly competitive markets are less well understood than the short-term effects (cf. Day et al., 2002; Joskow and Kahn, 2002).

Strategic behavior

The art of developing a system based upon reliability contracts or capacity subscriptions is to guard against new possibilities for strategic behavior. In general, the combination of a regulated volume of generation capacity and the fact that generation capacity cannot be expanded on short notice, means that there always is an opportunity for capacity withholding somewhere in the system. In the case of reliability contracts, the generating companies may manipulate the contract auction; in the case of capacity subscriptions, the generating companies may be able to artificially raise the prices of the subscriptions. The vulnerability of these capacity mechanisms must be tested, in a model and/or in practical tests, before they can be implemented with any confidence.

Other opportunities for manipulation may occur through exchanges with neighboring systems with different market models. Implementation of a capacity mechanism should lead to a larger reserve margin and lower prices when the system otherwise would have been under stress. During a regional shortage, there will be a temptation to sell to neighboring systems if these have higher prices. In this respect direct contracts between consumers and generating companies, as exist in a system of bilateral reliability contracts or capacity subscriptions, appear more robust than reliability contracts in which a central agency purchases options on behalf of the consumers.

Consumer behavior

An issue with capacity subscriptions is the behavior of consumers. If they do not understand their long-term interest of having sufficient capacity, they may under-contract for capacity, opening the perspective of an investment cycle. This cycle may dampen as consumers learn to purchase sufficient capacity but the learning curve may be costly. The likelihood of this scenario should be tested before implementation.

Assignment of responsibilities

It is the task of the system operator to preserve the operational reliability of the electricity system. To this end, he contracts system reserves (also called regulating power) with which he can correct imbalances between supply and demand in real-time. This obligation places system operators in energy-only markets in an awkward position with respect to the long term, as they do not have any means to influence the volume of available generation capacity. A survey of European countries shows that the responsibilities for generation adequacy generally are restricted to monitoring by the system operator or by a government agency (UCTE, 2002a). In some cases, there is a

planning requirement, however without a means to implement the plans. The actual provision of adequate generation resources is generally left to the market, in Europe.

If a choice is made to implement a capacity mechanism, responsibilities need to be assigned for:

- choosing the desired level of reliability of electricity service (except in the case of capacity subscriptions),
- operational decisions regarding the capacity mechanism, and
- how to monitor and enforce the system.

Except for capacity subscriptions, all capacity mechanisms have in common that the desired generation capacity margin is the same for all consumers: reliability is a public good. The choice of the level of generation adequacy, which determines system reliability, could in theory be made through a benefit-cost analysis. The marginal cost of providing a large capacity margin should equal the marginal social benefits of the resulting reduction in power interruptions. However, these calculations are difficult to make, in particular because the social cost of service interruptions is so difficult to determine. As a result, the level of reliability becomes a political choice, in which the cost of electricity is weighed against the perceived acceptability of occasional service interruptions.

Table 8-2: Responsibilities with respect to generation adequacy

	government or system operator	market	consumers
strategic reserve	• determination of the reserve margin • operation of the reserve capacity • monitoring and enforcement		
operating reserves	• determination of the reserve margin • operation of the reserve capacity • monitoring and enforcement		
capacity requirements	• determination of the reserve margin • monitoring and enforcement	• operation of the reserve capacity •	
reliability contract	• determination of the reserve margin • monitoring and enforcement	• operation of the reserve capacity	
capacity subscriptions	• monitoring and enforcement	• operation of the reserve capacity	• determination of the reserve margin

The second issue is who makes the operational decisions. This depends upon the capacity mechanism that has been chosen. In a centralized system such as operating reserves pricing or a strategic reserve, the system operator decides when to dispatch the reserve units. Capacity requirements and reliability contracts leave this decision to the market: they place the responsibility to provide a certain level of generating resources with the market. The same is true of capacity subscriptions.

Monitoring, finally, is a function that should be performed by an independent agent, so either a government body (such as the regulator) or the system operator are likely candidates. Table 8-2 shows an overview of the distribution of responsibilities under the different capacity mechanisms.

8.5 Conclusions

Interconnected systems, such as the continental European electricity markets, should jointly implement the same capacity mechanism. If they fail to do so, unilateral implementation of a capacity mechanism by some of the interconnected systems will distort trade and therefore reduce economic efficiency, while it is more difficult to devise an effective capacity mechanism for an open market. Moreover, a capacity mechanism that is implemented unilaterally would need to replaced when a regional solution is developed later in time. Long-term commitments that were engaged in under the first capacity mechanism could become stranded costs in the transition to the new, regional capacity mechanism.

In an open, decentralized system, reliability contracts become highly complex. A possible alternative for open, decentralized systems (like most European markets) is a mix of capacity requirements and reliability contracts, dubbed ‘bilateral reliability contracts’ in this chapter. A similarity with PJM’s system of capacity requirements is that the load-serving entities are required to purchase a certain volume of contracts; the similarity with the reliability contracts proposal is that these contracts are options for electricity, rather than capacity credits. Bilateral reliability contracts, however, may not be compatible with vertical integration of generating companies with retail companies. In the presence of real-time meters at every consumer, a financial version of capacity subscriptions would both reduce the implementation requirements and make the system robust against inter-system trade.

The choice of capacity mechanism depends upon the conditions. Is it to be implemented in a decentralized system or in an integrated system? Does the system rely upon imports, and if so, how is generation adequacy maintained in the exporting system? How much time is available before a shortage of generation capacity is expected? The above capacity mechanisms require time to be developed and implemented; if time is running out, a strategic reserve or operating reserves pricing may provide temporary relief. These capacity mechanisms are not robust, however, against regional shortages and appear less effective in stabilizing investment.

While the introduction of a capacity mechanism carries an inherent risk of design flaws

and new opportunities for manipulation by the market parties, capacity mechanisms may also provide additional benefits. For generating companies and consumers alike, more stable prices reduce risks. A number of capacity mechanisms also provide incentives for demand to exhibit a higher price-elasticity, which contributes directly to the overall economic efficiency of the electricity supply industry. Finally, to the degree that a capacity mechanism leads to a higher volume of available generation capacity and therefore fewer shortages, it also reduces the development of market power related to shortages. The possibility of imposing a maximum price in some capacity mechanisms further reduces the opportunity to abuse market power.

8.6 Recommendations for European markets

The current European policy of leaving generation adequacy to subsidiarity is undesirable. Only at a regional level is it possible to implement a capacity mechanism that is effective, efficient and robust. Due to the decentralized nature of most European electricity markets, it is much more difficult for individual countries to take effective measures. Considering the arguments provided in Chapter 5, a capacity mechanism should therefore be implemented by the EU. Preferably a single mechanism is implemented in as large a part of the interconnected system as possible. Peripheral systems with weak links to the main continental network may be allowed to choose a different capacity mechanism.

If implemented in a system with limited outside trade, capacity requirements would be an option that has been proven to work. This is the case in some European countries but more importantly it also is true for Europe as a whole. The fact that this capacity mechanism has been tried successfully makes this an attractive option. Both the central and the bilateral versions of reliability contracts, on the other hand, also appear to be effective while they promise to be more efficient. The lack of experience is the main disadvantage of these capacity mechanisms.

In the absence of EU policy to stabilize the volume of generation adequacy, individual member states may decide to implement a capacity mechanism. In these cases, inter-system trade will play a significant role in many cases. Capacity requirements do not appear effective in decentralized, open systems because there does not appear to be a way to 'recall' exports during shortages. This leaves decentralized, open electricity systems with a choice between the central and a bilateral variant of reliability contracts. Both provide a clear investment signal and incentives to generating companies to maximize their output during shortages, as well as to objectively estimate the volume of available generation capacity that they control.[51] A disadvantage of central reliability contracts in an open, decentralized system is their complexity; a disadvantage of the bilateral

[51] Note that in the USA, capacity requirements always are implemented in integrated systems, where the market operator also is the system operator.

reliability contracts is that this capacity mechanism appear less effective in the presence of vertical integration of generation and retail.

9 Coordination of generation investment with the network

This is the first of two chapters which discuss the issue of the coordination of the generation market with the networks in European electricity systems. The physical relations between electricity generators and networks complicate the economic goal of unbundling, which is a requirement for fair competition in the generation market. In principle, adequate financial incentives can be created to stimulate generators to coordinate their operational and investment decisions with the electricity network. However, for reasons of transparency and expediency, European countries have chosen for a relatively simple system based upon transmission tariffs. These create externalities which may lead to inefficient incentives. A number of policy dilemmas are the consequence. This chapter frames the issue and outlines policy options. The next chapter explores one of these options.

9.1 Introduction

In this chapter the relationship between electricity generation facilities and electricity networks is explored. Whereas the previous chapters focused on the quantity of generation capacity, now the relationships between the generation market and the electricity network will be considered, such as the physical location of generation units within an electricity network.[52] Before liberalization, network development and investment in generation facilities were coordinated in order to minimize overall cost, given certain reliability targets. Often, planning constraints limited the development of power lines, so the development of generation stock had to be adjusted to the physical possibilities of the electricity network. While of central planning of generation capacity is

[52] Technically, one cannot speak of 'the' electricity network, as an electricity system contains a number of networks of different voltages, linked to each other through transformers. We will use the term network to indicate the whole of all connected networks and supporting equipment that is under the control of a single system operator. The term location will refer to the connection point of a generator or load to the network, including the voltage level. The location of active generators and active loads determines the load flow pattern through a network.

anathema to liberalization, it does not mean that the goal of system-wide economic efficiency no longer exists. To the contrary, improving overall economic efficiency of the system was an important motivation for liberalization.[53] However, it appears that in some cases that goal has been overshadowed by the goal of introducing competition in as many areas as possible.

In the electricity sector, competition is limited to electricity generation, trade and delivery, while network operation remains a monopoly service. Unbundling – the separation of monopoly activities from competitive activities – is widely considered to be a necessary requirement for creating a level playing field for all competitors in the electricity industry (cf. Newbery, 2001; FERC 2002b; Directive 2003/54/EC). This chapter focuses on the particular case of European electricity systems. The analysis in this chapter applies only to electricity systems that are effectively unbundled, that is, in which the network companies have no economic interest in any party active in the electricity market. A consequence of unbundling is that the electricity system is no longer planned in an integral manner: the monopoly functions are regulated, while the competitive activities are free, within the usual limits that apply to businesses.

The technical reasons for coordination between generation and the network exist regardless of the economic model upon which the design of the electricity market is based. A lack of coordination of investment in generation capacity and in network capacity will likely increase system cost and may also reduce the quality of electricity service. To provide an example, in the course of time, generators may move away from consumers to locations where input costs are lower. The costs of network capacity expansion to accommodate such a move is not necessarily offset by the reduction in the cost of generation. A second issue is that network expansion typically takes much more time than the development of generation capacity and often is blocked by planning limitations. Inefficient locational decisions by generators may cause congestion and may eventually also lower the reliability of the system. In an unbundled system, competition provides an incentive for parties in the electricity market to be efficient, and the regulator may attempt to maximize the efficiency of the network companies that he regulates but their joint development is not necessarily efficient as well.

The presence of significant physical interdependencies between the network and generators needs to be reflected in the economic and institutional design of the sector. To maximize system efficiency, generating companies should receive incentives for optimizing their output and the location of new capacity within the constraints of the network. Similarly, network operators should receive incentives for optimal network operation and, especially, investment. While the cost of transmission networks is relatively low, the development of transmission networks often is physically constrained. This means that, in practice, the issue of coordination mainly involves providing incentives to generating companies for efficient operation and investment within the constraints of the network. Therefore the main question that is addressed in this chapter is to what extent it is necessary to provide generating companies with incentives to adjust

[53] See for instance the opening statements of EU Directive 96/92/EC; also the explanatory memorandum for the revision of this directive (EC, 2001c).

operating and investment decisions to the physical constraints of the electricity network, and how this can be done. A related second question is to what extent network operation and development can and should be adjusted to accommodate the generation market.

In principle, network tariffs are the obvious choice of instrument for providing generating companies with adequate incentives for the use of the network. However, the European choice for transmission tariffs that are fixed *ex ante* means that these tariffs do not reflect the real-time load flow conditions of the network, and therefore cannot provide efficient incentives for the operation and the development of the generation and network sectors. This means that other mechanisms need to be deployed to ensure that the transactions in the electricity market have a physical result that is feasible within the constraints of the electricity network. This chapter explores the tension between the use of fixed network tariffs and the need for coordination of generation with the network.

To provide a definitive answer about the costs of insufficient coordination between generation and network, one would need to use a quantitative model of a network. Empirical data on the long-term costs of insufficient coordination is limited, due to the relatively short history since liberalization, and difficult to obtain. As we are only beginning to observe some of the issues that are the subject of this chapter in practice, models could help by forecasting long-term developments. However, the development of detailed network models is outside the scope of this research project. This chapter uses a qualitative approach to structure the potential problems that may arise as a result of insufficient coordination and to assess possible solutions. The focus is on areas where inadequate incentives may create serious inefficiencies, and possible remedies. Thus, this chapter's main contribution is to present a systematic problem analysis, from which possible solution paths can be derived. Chapter 10 will further analyze one set of possible solutions, namely congestion management methods for unbundled networks with fixed transmission tariffs.

The argument that operation of and investment in generation capacity should be coordinated with the network also applies to loads. In an efficient electricity market, consumers are also confronted with the costs of the consequences of their decisions. This issue is not included in the analysis because, firstly, changes by small consumers (changes in location and shifts in consumption patterns) largely cancel each other out and secondly, because loads can be modeled analogously to generation, so the same types of instruments are available to influence the short and long-term behavior of consumers as for generating companies.

Reading guide

The next section starts with elaborating on the analytic framework that was presented in Chapter 2. Section 9.3 presents a brief recapitulation of the generic policy goals for the sector and develops them with respect to coordination of generation and the network. Section 9.4 reviews the physical relationships between generation facilities and the network. Next, The perspectives of the main groups of actors are reviewed in Section 9.5. Section 9.6 contains the central part of this chapter. In designing a system with fixed transmission tariffs, a number of dilemmas emerge when a comparison is made of the

policy goals and the physical relationships. Section 9.7 reviews the options for improving coordination. The last section before the conclusion, Section 9.8 discusses the need for a paradigm shift in the design of the European markets.

9.2 Analytic framework

The conceptual framework that was developed in Chapter 3 provides the analytic framework for this chapter. For the sake of convenience, Figure 3.8 is repeated in Figure 9.1. The figure reflects one of the essential features of liberalized systems, namely that system operation and development are not guided by a planning agency but by a number of actors. Coordination is therefore to be achieved through the proper structuring of the relations between the actors.

The relationships between generators and the network are determined by the laws that govern the sector, the decisions of regulatory authorities, the physical and technical conditions in the sector and – last but not least – by the actors within the system, through their conduct and the contracts into which they enter. In the terms of Figure 9.1, the economic relationships between the producers and the network managers need to reflect the physical relationships between the generation facilities and the networks that they control. If the structure of the technical subsystem is insufficiently reflected by the structure of the economic subsystem, external costs and benefits may develop. They may cause the system equilibrium to deviate from the economic optimum.

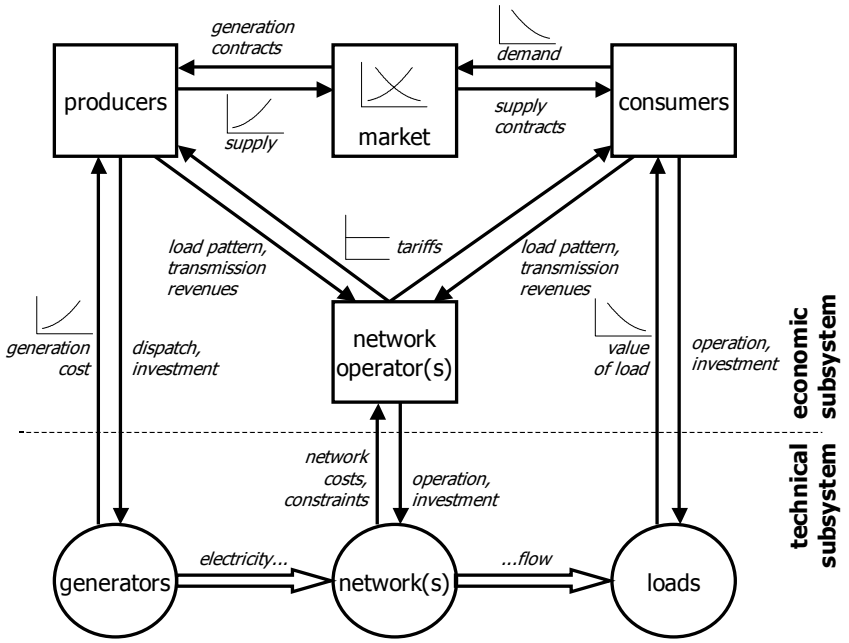


Figure 9.1: Conceptual framework (See Chapter 3)

The effect of introducing competition in the generator market, regulating the networks and certain other services as a monopoly, and applying different, specifically tailored market mechanisms for some other services, is a fragmentation of the economic structure of the electricity system, whereas technically, the system is as strongly integrated as before. A liberalized electricity system in fact is a hybrid between a market – or actually a set of markets – centered around the production and delivery of electricity, and a number of monopoly functions, centered around the network. In theory, this fragmentation should be managed by the introduction of appropriate economic incentives to the actors. Where competition is not possible, the economic incentives should be structured so each agent has the correct incentives for contributing to the economic efficiency of the overall system.

Focus on Europe

This chapter focuses on decentralized electricity systems because this model is dominant in Europe. (The most notable exception is NordPool.) In this model, transmission tariffs are calculated separately from the electricity price and stated *ex ante*. They are based on average network costs and are not distance-related. They may, however, include locational variations in access charges (Regulation (EC) No 1228/2003). The European system of fixed transmission tariffs is fundamentally different from the system of locational marginal pricing (also called nodal pricing), which is used in the PJM system in the USA, and which is the FERC's preferred model (FERC, 2002b). The idea behind the European model is that the networks should facilitate the development of the electricity market and that therefore network tariffs should be simple and transparent (Regulation (EC) No 1228/2003).

Locational marginal pricing is a way to provide efficient locational incentives to generators (Hogan, 1992; Stoft, 2002). In this system, which requires an integrated electricity system, generators do not deliver electricity to specific customers but to the market operator. He purchases electricity from each generator at its 'node', its connection point to the main network, and delivers electricity to consumers at their own nodes. Generators and consumers at each node submit supply and demand bids, which the market operator uses to calculate the equilibrium price and volume of supply and demand at each node. The market operator incorporates the operational costs to the network of injecting and withdrawing at each node in the price of electricity at that node. Within the physical constraints of the network, the network operator minimizes the cost of electricity production. This way he creates a system-wide operational optimum.

In addition to the desire to have a simpler and more transparent system, a likely second reason why locational marginal pricing is not used in Europe is that implementation does not appear feasible in the foreseeable future. As it is applied in the USA, locational marginal pricing requires a high degree of standardization of market rules and a single system operator who also is market operator (Wu et al., 1996). This alone is a goal that appears unreachable for Europe in the near term. The countries which constitute the common electricity market each have their own set of market rules that differ widely from each other. Some countries also have progressed farther along the path of liberalization than others. A final argument against locational marginal pricing in Europe

appears to be that the geographic variations in prices could be considered inequitable (Crampes and Laffont, 2001).

A different kind of objection against locational marginal pricing is that it resembles the former centrally operated systems, as it concentrates much power in the hands of the system operator (Wu et al., 1996; Rosenberg, 2000). European countries generally have chosen for a decentralized system with network charges that are fixed *ex ante*.[54] The latter leaves more freedom to agents in the sector to devise their own solutions, as is for instance witnessed by the development of a number of private power exchanges. *Ex ante* fixed tariffs are intended to facilitate the market by making the system simple and transparent. However, this chapter will argue that this is at least partly an illusion due to the need for different kinds of corrective measures.

9.3 Policy goals

The general policy goals for the electricity supply industry are to provide power reliably, efficiently and in a sustainable manner (Directive 96/92/EC). The environmental sustainability of the electricity supply industry is mainly determined by the choice of generation technology and the use of primary energy sources, which are not at issue here. The reliability of service only enters into the analysis for as far as it is impacted by the relationship between the generation market and the network.

The main instrument to achieve economic efficiency is competition. In addition to generation, it may be possible to arrange a number of other services within the electricity system competitively (Hakvoort, 2000). For example, successful markets have been established for system balancing services. Congestion may also be managed through market-based mechanisms, such as auctioning of network capacity. For ancillary services, such as reactive power management, similar competitive mechanisms have been proposed (Kirsch and Singh, 1995). As the system description in Chapter 3 also showed, a variety of different markets within the electricity system may develop. A trade-off will need to be made between the inevitable imperfections of each market and the alternative, which is to include it with the network monopoly. The introduction of many sub-markets for different parts of the electricity system increases the complexity of the system, as these markets are all related to each other.

To improve the economic efficiency of the network, competition between multiple networks will not be considered an option in this chapter. The electricity network is widely assumed to constitute a natural monopoly because the costs of multiplication, in order to create competition between different networks, exceed the potential benefits from competition.[55] Even though the status of the network as a monopoly does not change, liberalization necessitates a change in network regulation. Unbundling of network and competitive functions is required to create equal conditions for competing generation companies. As part of the restructuring process, in some systems policy

[54] The main exception is Norway, in which market splitting (a form of zonal pricing) is used.
[55] For an alternate view, see Künneke (1999).

makers have also attempted to improve the economic efficiency of the networks by changing the regulatory incentives. In the spirit of liberalization, they have replaced cost-based network tariff regulations with incentive-based regulations (cf. Ajodhia, 2002a). Rather than providing a fixed rate of return, they allow the surpluses of network managers to increase with their economic efficiency. This is, however, a different issue, which will not be discussed here.

Summarizing, the goal of improved efficiency is to be obtained through the introduction of financial incentives where possible. Competition is possible in generation and supply (retail) of electricity. Market mechanisms may also be used for at least some of the system operation activities (system balancing, voltage control, congestion management). Network management, on the other hand, is a widely accepted natural monopoly. To stimulate efficiency, incentive regulation may be applied. In order to ensure a level playing field in the electricity market, competitive functions must be fully unbundled from monopoly functions. For the sake of system-wide economic efficiency, the relations between the generation market and the network monopoly need to be structured with efficient incentives.

9.4 The relations between electricity networks and generators

9.4.1 Introduction

To analyze the effects of unbundling, first the relations between electricity generators and the networks to which they are connected need to be understood. These relations have consequences for the economic and institutional design of the sector, as will be seen in Section 9.6. Generators impact electricity networks in multiple ways. They do not only contribute to the demand for network capacity, they also play a role in network operation. Vice versa, the network constrains the market activities of generators, for instance through its geographical structure and its capacity. In this section the main relations between generators and the network will be discussed, starting with operational aspects of load flow and voltage control and moving towards investment issues.

9.4.2 Load flow

The primary function of the electricity network is to transport electric energy from generators to loads. The network costs of a transaction – a sale of electricity from a generator to a load – vary with the load flow. Given a certain network and a certain combination of active generators and loads, the load flow through that network can be calculated with the laws of physics. Essentially, if there are multiple parallel paths along which load may flow, the electricity uses all possible paths according to their relative impedance (resistance). The higher the impedance of a line, the more electricity will flow through parallel lines, if they exist. When electricity is transmitted from a generator to a load, part of the electricity may take quite long ‘detours’, using network connections far away from the shortest path. This phenomenon is often referred to as loop flow but

parallel flow is a more accurate description. An important and perhaps counter-intuitive conclusion is that the flow through a line is not determined solely by its capacity. Consequently, when there are multiple parallel paths, it is not at all a given that the available network capacity is maximized. To the contrary, the load flow may be such that certain lines are overloaded while parallel lines are not used to their maximum capacity.

A peculiar and significant characteristic of electricity networks is that there are few instruments for their operation other than the adjustment of generator output. As a result, in practice the load flow is largely determined by the activities of generators and loads. The general consumption pattern is fairly static, as changes among the many loads largely cancel each other out. Therefore the main variable that determines the flow through an electricity network in the short term is the location of active generators. The market determines how much each generating company may produce; the companies themselves decide which units they operate. Thus, the larger companies have a significant impact upon the load flow through a network, as they can shift production between generation units at different locations. In the vertically integrated utilities of the past, operational control of the network was integrated with the dispatch of generators, and system development was also planned from an integrated perspective. In an unbundled system, the network operator needs to contract with generators for their services, such as voltage control and congestion management.

The main operational costs of networks are energy losses and congestion. Energy losses are determined by the load flow pattern, as a result of which it is difficult to attribute them to individual transactions. Network energy losses increase with the square of the current through a power line, which means that a doubling of the energy flow (at a given voltage) leads to a quadrupling of energy losses. Energy losses can be reduced by using higher voltages, which is why transmission lines use high voltages.

Congestion occurs when the combination of all market transactions causes a load flow pattern that exceeds the capacity of the network. Physically, this is an untenable situation, as overloading damages power lines. Therefore, electricity transactions need to be notified in advance to the network operator, who calculates the load flow that results from these transactions and takes measures if congestion is foreseen. An interesting aspect is that electricity flows in opposing directions cancel each other out and thereby reduce energy losses and congestion. Doubling an electricity flow in the same direction, on the other hand, more than doubles the associated energy losses and, of course, contributes to congestion. As a result, it is impossible to predict either the energy losses or the congestion costs of any given transaction without knowing the details of all other transactions that take place at the same time.

Congestion management methods return a certain measure of control of the load flow to the network manager. Some of them allow the network manager to intervene directly in generation output, while others work indirectly, through financial incentives to generating companies. See Chapter 10 for an analysis of congestion management methods. The congestion management methods also differ with respect to the incentives they provide to the network managers and to the generation companies. Apart from the choice and details of the congestion management method, network managers may influence the occurrence

of congestion in the short term through the way they calculate network capacity and through their choice of safety margins. In the long term, network capacity additions may relieve congestion.

9.4.3 Voltage control

Electricity must be delivered to customers within certain technical standards. It is a peculiar characteristic of the product electricity that its quality is determined by the system as a whole, rather than by the producer of the product. Most important, for our purposes, is the control of the voltage level. The voltage level is impacted by the load flow. Changes in the load flow impact the voltage differently in the different parts of the network. Consequently, voltage control requires active local intervention. It is the network manager's task to continually maintain the voltage of each part of the network within prescribed limits.

The voltage level is maintained through the production or absorption of reactive power. This is a function that generators can provide at little extra cost, as it can be provided as a by-product of regular power. Reactive power also be managed through the use of capacitors but these need to be purchased specifically for this purpose.

9.4.4 System development

Investment in generating facilities and in network capacity need to be coordinated to ensure that the network has sufficient capacity and to maintain the operational stability of the network. Physically, this relationship is quite visible, as each generator requires a physical connection to the network. A network connection is defined as the link between the customer and the nearest network node. A node can be defined as a point at which power lines connect to each other or, via a transformer, to a network of a different voltage. If a large generator or load is connected, or if there are many new connections in one area, the capacity of the network may need to be increased to accommodate the increase in energy flows.

Generators can be charged for access to the network, both for the cost of the connection to the network and for the necessary upgrades in the network itself. These charges may consist of a one-time connection fee and a returning access charge. They are typically related to the capacity of the connection and the voltage level that they are connected to.

9.4.5 Facilitating competition

Prior to liberalization, the challenge of system development was limited to supplying every consumer with electricity. The utility manager could use both generation and network investment to meet this goal. In a liberalized market, consumers should be able to purchase electricity from their suppliers of choice. This means that the network should not only be tailored to one, technically optimal dispatch pattern of generation facilities but that it should be able to accommodate a variety of generation scenarios. It may also mean that network capacity needs to be improved merely for the purpose of competition, for instance if network constraints cause a generating company to have market power in a

part of the network. This generating company may have enough units so the reliability of service is sufficiently ensured but his market power may still be sufficient reason for investment in network capacity so other generating companies also are able to serve the customers in that area (Borenstein et al, 2000; Hakvoort and De Vries, 2002).

9.4.6 Overview

Figure 9.2, which is based upon the left part of Figure 3.8, depicts the relationships that are described in this section. Generator dispatch determines not only the load flow but also influences the availability of reactive power management services for the purpose of voltage control. From the load flow pattern, the network managers determine the need for congestion management. This arrow is curved, as some congestion management methods use financial incentives to generating companies, while others provide the network operators with the authority to intervene directly in the dispatch of generators. The presence of congestion, if it is persistent, may also impact generators' locational decisions, depending upon the incentives provided by the congestion management method in place. (See also Chapter 10.) The locational decisions of generators, finally, determine the need for network connection capacity and influence network managers expectations for future demand for network capacity. In addition, there may be a need to expand network capacity for the purpose of creating more opportunities for competition, as is indicated by the arrow from network capacity to the market.

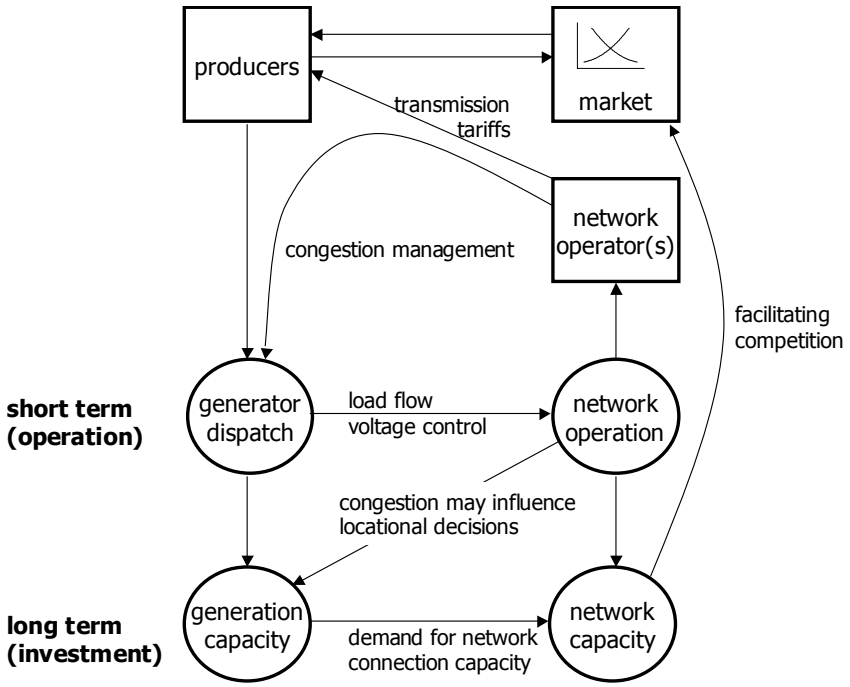


Figure 9.2: The relations between generation and the networks

9.5 Actor perspectives

9.5.1 The perspective of generating companies

Liberalization has freed generating companies from many of their previous responsibilities. They are not accountable for maintaining the balance between supply and demand, nor are they responsible for aspects of network operation such as voltage control. In the long term, they do not bear responsibility for generation adequacy. Generation companies may choose to contract these services to the TSO or distribution network providers but are not necessarily obligated to do so. Only when the stability of the system is threatened does the TSO typically have the authority to intervene directly in generator activities, but he needs to reimburse these generators for any costs made. This is not to say that liberalization has simplified the business of electricity generation: the increased uncertainty and risks of operating in a market bring about their own challenges.

Generating companies' interests depend upon the situation. On the one hand, it is in their interest that the network has sufficient capacity to transmit the electricity they produce to their customers in a reliable fashion. Depending on the type of congestion management method that is being used, congestion may form a significant market risk for generating companies. If one of the congestion pricing methods is used, the price for transmitting electricity to his customer may increase, effectively reducing the value of the electricity that the generation company produces.[56]

On the other hand may generators in a constrained area benefit from the existence of congestion. Within the constrained area the marginal cost of generation may be higher than outside, increasing the competitive market price. In addition, when transmission options are limited, generators may develop local market power. Thus, the interest of generating companies depends on whether they sell into a congested area, or whether they are located within a congested area, and upon the congestion management method in place. In any case, generating companies have a particular interest in the regulation of the network.

9.5.2 The network managers' point of view

Liberalization has placed the networks in a peculiar position. Network managers now face larger risks and their function has expanded somewhat, compared to the situation before liberalization, while their control options have been reduced. In some cases, the regulatory framework has also changed from cost-based regulation towards incentive regulation. Thus, the demands upon network managers have increased, while their means have been reduced.

[56] However, when redispatching or counter trading are used for congestion management (see Chapter 10), the position of generators may be different. In this case, they are allowed to trade as if there were no congestion, and some generators are reimbursed for helping relieve the congestion. These systems appear vulnerable to manipulation by the generating companies, as Enron proved (Behr 2002).

Increased risk

The increased uncertainty which network managers face is a consequence of the unbundling of networks and generation: while a combination of investment in the network and in generation is needed to meet demand, investment in the two is no longer necessarily coordinated in an unbundled system. Generator output is the main control variable for network operation. Unbundling has removed this variable from the control of the network managers. At the operational level, this loss is partly compensated through congestion management methods, which provide a way to influence the dispatch of generation. In the long term, however, network managers appear to have insufficient tools to coordinate the development of their network with the investment decisions of generation companies. As system conditions change, for instance due to an increase in demand or changes in the location of active generators, electricity networks are expected to continue to provide adequate links between generators and consumers of electricity. Not knowing future generator behavior, however, complicates network investment decisions. Network development takes longer and may be slowed down further in the permitting process. Similar uncertainties exist with respect to the availability of reactive power management services, which generators traditionally provide.

New role: facilitating the market

As operators of the back-bone of the electricity system, network managers also have a function of facilitating the electricity market: the network is expected to accommodate the flows that result from the market transactions. Network managers have no intrinsic interest in investing for the purpose of stimulating competition in the electricity market. If the networks are properly unbundled from the electricity market, the network managers should be indifferent to the competitive dynamics of the generation market. Investments for the sake of improving competition do not lead to higher overall usage of the network, so network managers will not recover them through transmission tariffs. Therefore network managers need a special incentive or obligation to make this type of investment.

Change in the regulation of network tariffs

Liberalization necessitates a re-regulation of the networks in order to arrange such issues as third party access. However, the drive for economic efficiency also extends to the networks, and in some cases has led some policy makers to change the regulation of network tariffs. Traditionally, network tariffs were often determined by some form of rate-of-return regulation, which allowed network managers a more or less fixed return on their investment. While this prevented excessive monopoly rents, it did not provide network managers with an incentive to be economically efficient. In the spirit of liberalization – improving economic efficiency through financial incentives – a new method of network regulation was introduced.

Incentive regulation is an attempt to mimic the effects of competition upon a network company (Pfeifenberger and Tye, 1995). By fixing the tariffs per unit of electricity transmitted, network managers become price-takers, just like competitive firms. Rather than being able to recover all their expenses, now they have a fixed budget within which they need to operate. The trend of declining prices in competitive markets is copied by

gradually reducing the maximum tariffs, which forces the management of the networks to become ever more efficient. This system is usually limited to expenses that the network managers can control, or even further limited to only operational expenses. Thus, it may be combined with rate of return type of rules, for instance the possibility of raising the tariff to recover specific kinds of investment. The fact that certain costs are not recovered with the same degree of certainty as under rate-of return regulation should make network managers more prudent with investments, perhaps even risk-averse.

9.5.3 The interest of consumers

The general goal for an electricity system is to provide electricity reliably, at a reasonable cost and with a limited environmental impact (cf. DTe, 1999; Directive 96/92/EC). The main objective of liberalization is to introduce competition where possible, as this should improve the efficiency of the affected activities. This way, a liberalized electricity system should be able to achieve the same policy goals as a monopolistic system but at a lower cost. Assuming that these policy goals were established in a democratic manner, these will also be taken as the main objectives of consumers in this analysis.

An important issue is the question of the optimal ratio between the cost and the performance of the electricity system. Historically, the balance between cost and performance often was not made consciously, which led to an arbitrary result. Reliability was maximized within a certain budget that appeared reasonable, based on experience. With the increased focus on economic efficiency, the question emerges which cost reductions are efficient and which ones create a disproportionate loss of quality of service. The theoretically ideal way to find the economically efficient trade-off between cost and quality is the market mechanism but this solution is excluded by the network monopoly. Another characteristic of networks is that groups of customers receive the same quality of service, even though individual preferences may differ. A partial solution for the differences in individual preferences is provided by capacity subscriptions, which allow consumers to choose their level of reliability of electricity generator (see Section 6.7). However, as long as there is a network monopoly, the influence of the network upon the quality of service will be uniform among all consumers who use the same part of the network. The issue of network quality is further developed in a related research project (Ajodhia, 2000a and 2000b).

9.6 Five market design dilemmas

From a comparison of the policy goals, the actor perspectives and the physical relationships between generators and electricity networks, a number of policy issues emerge for decentralized electricity systems. In each case, the regulatory structure does not reflect the physical characteristics of the sector. The resulting policy questions can be framed as five dilemmas. These dilemmas reveal tensions between network management and the generation market, which stem from the physical characteristics of the network and the different paradigms within which the generation market and the network exist. Two areas are addressed: the question how to structure incentives to the generating companies and network development in response to developments in the generation

market.

9.6.1 Load flow

Unbundling raises a number of challenges with respect to system-wide efficiency (Haubrich et al., 1999). Section 9.4 explained that electricity transactions may create two types of operational costs for the network: energy losses and congestion costs. Both are impossible to predict, unless one already knows all other transactions that will take place at that moment. If it is impossible to determine the actual network cost of a specific transaction *a priori*, how can transmission services be priced so they provide efficient incentives? [57]

As explained in Section 9.2, most European countries have decentralized electricity systems with fixed network tariffs are determined separately from the market price.[58] This is a consequence of the choice for a decentralized electricity system. As the transmission system operator and the market operator are different agents, network management is performed independently from the electricity market. This creates a significant dilemma. As the network costs of each transaction depend upon the joint effect of all other transactions, they can not be determined *ex ante*. However, billing network users with *ex post* tariffs would appear unreasonable and, worse, constitute a significant barrier to trade due to their unpredictability: who would want to engage in an electricity transaction without knowing the associated costs of network use in advance?[59] A choice for standardized tariffs appears inevitable. However, tariffs that are fixed *ex ante*, such as a price per unit of electricity transported that is fixed and independent from distance, route, or time, create significant network externalities. The fact that these externalities are substantial can be concluded from the fact that the operational network costs of some transactions are zero (when they go against the prevailing load flow), while others have costs that are much higher than the average cost, for instance if they stress the available capacity so they cause high energy losses and perhaps also congestion.

> *Dilemma 9.1: Value-reflective network charges fluctuate unpredictably and are therefore unacceptable to network users, unless they are integrated in the energy market. By definition, this is not possible in a decentralized system, but network charges that are not value-reflective create significant externalities.*

The use of fixed tariffs to recover network costs implies that network costs are socialized.

[57] The only currently available solution is nodal pricing. In this system, all demand and supply bids are made to a pool operator, who establishes a separate electricity price for each node in the network. By varying the nodal prices, the pool operator determines which generators are in merit, and so influences the load flow. Through this mechanism, the least-cost dispatch can be found that does not create congestion. Transmission losses can be included in the optimization algorithm. See for instance Hogan (1992), Oren et al. (1995), Stoft (2002).

[58] The NordPool market in Scandinavia forms an exception. Here a system of market splitting is used, which resembles nodal pricing with the exception that only a small number of large nodes is used.

[59] Nodal pricing solves this problem by combining energy and network costs into single bids in which the system operator is the counter party for both generators and consumers.

This means that an important opportunity to provide generators with efficient incentives for the use of the networks is forfeited. For instance, generating companies will not take energy losses into consideration when deciding which generators to operate. Consumers therefore also do not receive an incentive to minimize network losses or congestion through their selection of electricity provider. For system operation, the increased energy losses have no great consequences. They do represent a loss of economic efficiency and an increase in environmental cost. Congestion, on the other hand, must be addressed. Congestion management methods (Chapter 10) can be considered *ad hoc* remedies for the most extreme side-effects of transmission tariffs that are not value-reflective.

9.6.2 Voltage control

Traditionally, the voltage in each part of the network is kept within specified limits through the management of reactive power, which generators can produce and absorb. In systems in which the networks are effectively unbundled from the generation market, network managers may find themselves in the awkward position that they are dependent upon generators for the operation of their networks. In a dense network, where there are many generators relative to the size of the network, it may be possible to create a market for reactive power (Hogan, 1993; Murray, 1998). Then, competition between generators should ensure that the network manager pays a reasonable price for their services. However, due to the local nature of reactive power, the likelihood of the development of a local monopoly is high. There simply may not be sufficient generators active in a specific part of the network to create effective competition in the provision of reactive power. Moreover, the legacy of the former regional monopolies increases the likelihood of a local monopoly over the provision of reactive power.

Another difficulty with relying upon generators is that reactive power management is a by-product of real power. As market contracts usually are short (typically a year or less), the future availability of reactive power services from specific generators is uncertain. The network manager may find that for the provision of ancillary services he is dependent upon the commercial success of certain generators, which conflicts with the policy goal of an independent network manager. Long-term contracts between the network manager and the providers of ancillary services may be a solution, as they give the network manager time to create alternatives such as installing capacitors if a generator tries to abuse his monopoly power (Kirsch and Singh, 1995). However, generators may be reluctant to enter into reactive power contracts with a longer duration than their regular power contracts.

It appears unattractive for network managers to depend upon generating companies for the provision of reactive power. An alternative is to install capacitors, which can also be used to control the voltage, at various locations in the network. As this is more costly than the provision of reactive power by active generators, it appears to violate the goal of economic efficiency but perhaps a trade-off is necessary. This brings us to the second dilemma:

> *Dilemma 9.2: For the provision of reactive power, network managers have the choice of relying on local generators, which is risky, or investing in capacitors, which is more expensive.*

The issue can be divided into two elements: the question of regulating monopoly power over the provision of ancillary services, if this exists, and the problem that the network manager is dependent upon generator behavior, which he cannot predict. Even in a competitive market, the future availability of generators' services to the network may be uncertain. The monopoly issue is not a new one but the second issue is a fundamental problem in a deregulated system. Perhaps it simply is one of the costs of liberalization that certain investments need to be made in the networks in order to make the network managers independent from the market parties. From conversations with network managers in the Netherlands, it appears that network managers are willing to pay the higher cost of capacitors in order to be independent from the generation market. By installing capacitors, network managers purchase independence from the generators and reliability. Kahn and Baldick (1994) suggest that the use of reactive power compensation equipment would be limited inexpensive and has as a benefit that the potential of generators to manage reactive power would remain available for emergency situations.

9.6.3 Locational incentives to generators

While unbundling complicates network operation, perhaps the most important challenge is to coordinate the long term development of generation and the network. The next two dilemmas are related to the long-term coordination of investment in generation capacity (and withdrawal of existing capacity) and the development of the network. Generators' decisions regarding new capacity or withdrawal of existing capacity may create disproportionately high network costs or even reduce system reliability.

Fixed transmission charges are aimed primarily at recovering network costs and do not offer an opportunity for coordination of generation investment with network development. Congestion management methods provide operational solutions to load flow problems and may provide efficient long-term incentives, albeit not with a high geographic resolution. A third option to guide generator investment decisions is through the incentives provided by connection charges, which will be discussed now. Perhaps they can be used to compensate the shortcomings of the first two, in terms of long-term efficient signals.

The establishment of efficient network access charges is complicated by similar problems as were encountered with transmission tariffs. Again, cost-reflective access charges appear infeasible, due to at least three mutually related obstacles: the question of deep versus shallow connection costs, the first mover problem and the lumpiness of network investment.

Generators and loads need to pay for their connections to the network. The costs of establishing a link to the nearest network connection point are called the shallow connection costs. New generation capacity or a large load increase, however, may also require expansion of the main network links, called the deep connection costs. These

costs usually are not easily allocated to specific users for several reasons. The first reason is the basic problem with allocating transmission costs: the need for network capacity expansion depends upon the use of the network by other generators and loads. As the load flow pattern is the result of the combination of all generation and all consumption, it is difficult to allocate the costs of network operation and of capacity improvements of the main infrastructure to specific users. This is all the more difficult as the load flow is affected by market conditions that may change rapidly.

The second issue with charging deep connection costs is the first mover problem. When a generator connects to the network at a certain point, this may trigger the need for network investments that benefit others also. Who is to pay for these expansions? How to handle time differences, for instance if the second party to benefit from the network expansion arrives later? This issue is closely related to the third issue: the fact that cost-effective capacity expansion only is possible in sizeable 'lumps', not in marginal increases (Turvey, 2002). As a result, one generator locating in a specific place may not trigger a need for network improvements at all, while the next generator of the same size may be the cause for capacity expansion far in excess of its own needs.

Due to these complications, network connection charges usually are based upon the shallow connection costs, while the deep connection costs are socialized. They can be included in the connection charge or as part of the transmission tariff. The resulting lack of appropriate economic incentives raises the issue of how to develop the optimal combination of generation and network capacity in each area in the network. Investment in generation capacity reduces network cost. However, in an unbundled system it may be difficult for a network operator to stimulate local development of generation. As loads do not pay cost-reflective network charges for the same reasons as generators, local generators appear more expensive to them than network improvements. Therefore they are likely to request network improvements even if an increase in generation were less costly for the system as a whole. Vice versa, generators may choose to locate in places that cause disproportionately high network costs because they are not confronted with them.

> *Dilemma 9.3: Unbundling prohibits integrated planning of investment in generation and network capacity by a single firm, while the varying nature of network costs hamper the creation of efficient incentives through fixed transmission tariffs and network access charges.*

A pragmatic solution may consist of varying access charges in order to influence the locational decisions of generators, including the voltage level of their network connection. If they are used this way, the access charges are used as proxy incentives. That is, the incentives given through the access fees are not a reflection of the actual connection costs but an attempt to compensate for the lack of incentives provided by the transmission tariffs. They can be used to correct both the shortcomings of the transmission charges (network operating costs) and of the network access charges (network capital costs). An example is the use of lower access charges in the south of England, in which demand exceeded the available generation capacity, while there was excess capacity in the north.

Because access charges that are used in this way serve to reduce transmission and congestion costs, they need to be recalibrated periodically, dependent upon the effect they have upon generator behavior. As congestion pricing can also be used to provide locational incentives, there are two instruments available to guide efficient operation and development of the system. Section 9.7 elaborates on these options.

9.6.4 Network development

The long-term coordination issue is further complicated by the long life-cycle of network components and the long lead time between the design of network improvements and their realization, relative to changes in the generation market (cf. Budhraja, 2003). The speed with which supply and demand change can be much greater than the speed with which network capacity can be adjusted. Power plants, especially the small gas-powered combined heat and power plants, can be constructed in much less time than most network improvements can be realized. Power plants can be taken out of service much faster. Even if it were possible to determine an optimal network design, it would therefore probably not be realized before system conditions had changed significantly. Network development inevitably lags behind market developments.

The fact that network expansion typically takes more time to realize than investment in generating facilities and, especially, the decommissioning of generation facilities places network planners for some difficult questions. The uncertainty regarding the future location of generators (and to a lesser degree loads) substantially increases the risks associated with network planning. This risk is further increased when network revenues are regulated with an incentive-based system, as for instance is the case in the Netherlands and the U.K. because it does not allow network companies to recover their costs automatically (Pfeifenberger and Tye, 1995).

Network development has always been complicated by the length and the difficulties of the permit process and the decades-long life cycle of network components. Prior to liberalization, the integrated planning process dealt with this issue. Generation facilities simply were not constructed in places where the network did not have sufficient capacity to transmit their power; on the other hand it was possible to make up for network capacity constraints by placing generators strategically. Moreover, integrated planning ensured that generation facilities were not closed down unless there was sufficient capacity elsewhere in the system, both generation and network capacity, to replace them. The planning process could not prevent the system from being slow, relative to changes in demand, but it could ensure that generation and network capacity were adjusted to each other.

The reduction of coordination possibilities amplifies the difficulties created by the long lead time for network development. Network managers are presented with a choice between investing in anticipation of market developments, which is risky, or letting investment lag behind market developments, which may result in a reduction of system reliability. Were network development more versatile, insufficient coordination could be compensated by speedy adjustments to the network.

> *Dilemma 9.4: Due to the long lead time for network investment and the long life cycle of networks, network investment in anticipation of market developments is risky and therefore involves higher average cost; however, reacting to changes in market demand means being substantially too late, which also creates high social costs.*

Solving the dilemma is complicated by the fact that the higher costs of an anticipatory investment strategy accrue to the network companies, while the costs of a reactive strategy are mostly for the generating companies and society.

9.6.5 Facilitating competition

In a liberalized electricity system, it may be beneficial to invest more in networks than would appear rational in a centrally planned system, even if the coordination with the generation market would be optimal. Generally, the larger the network capacity, the smaller the likelihood that generators have local market power. The paradoxical situation develops that additional line capacity may induce generators to more efficient behavior by creating more competition, while it does not necessarily lead to an increase in the use of the network. The mere presence of sufficient capacity to allow competition to develop is enough (Borenstein et al., 2000). Thus, apparently useless network capacity can provide a benefit to society through a reduction of market power. Network managers cannot recover investment in such network capacity through regular network tariffs, as the benefits manifest themselves solely in the form of improved competition, which does not necessarily increase the load flow.

> *Dilemma 9.5: To what degree should network capacity be increased to enhance competition when market conditions, and hence the benefits of such capacity improvements, may change on a much shorter time scale than these network improvements take place?*

This dilemma actually is a trade-off between the cost of additional network investment and the social benefit from increased competition in the generation market. It may not be efficient to expand the network to the extent that every consumer has a choice of a number of suppliers. For instance in remote areas, it may be preferable to simply allow local generation monopolies to emerge. However, they should be recognized and regulated, even if the monopoly only exists part of the time, for instance only during peak demand periods. In principle they can be regulated with traditional monopoly regulation methods (e.g. Crew and Kleindorfer, 1985). While a static benefit-cost analysis of additional network capacity for the purpose of stimulating competition may not be so difficult, the long lead time and life cycle of network capacity enhancements (the subject of the fifth dilemma), also are significant obstacles here. Not knowing whether the market will develop more competition of itself in the future, for instance in the form of distributed generation, the decision to embark on major network expansion projects for the sake of stimulating competition may indeed pose a dilemma.

Table 9-1: Overview of the dilemmas and their causes

section/ issue	physical aspect	constraints	ideal solution	dilemma	practical options
9.6.1: load flow	managing available network capacity, energy losses	unbundling: no direct control of generation	value-reflective network charges	9.1. the need for unbundling vs. the need to coordinate generation operation with the network	• flat network tariffs, combined with corrective instruments such as congestion management methods and permit requirements • nodal pricing
9.6.2: voltage control	reactive power management		many suppliers	9.2. local quality of product easily leads to local market power; technical solutions more expensive	• surveillance against abuse of market power • regulation • phase-shifting equipment
9.6.3: long-term incentives to generators	network connection, network capacity		value-reflective network charges	9.3. the need for unbundling vs. the need for coordination of generation investment with the network	• fixed tariffs, proxy charges, permits • nodal pricing?
9.6.4: network development	long lead time for investments, long life cycle of components		coordination of network and generator planning	9.4. invest too late or risk non-recoverable investments	more flexible network structures?
9.6.5: facilitating competition	load flow, number of generators that can serve each customer	network managers do not benefit from competition	investments based upon social cost-benefit analyses	9.5. invest in network capacity for the purpose of stimulating competition?	government performs cost-benefit analysis, decides when to allow network managers extra revenues to recover investments for the purpose of improving competition

9.6.6 Overview

Table 9-1 presents an overview of the dilemmas and how they relate to the policy goals and the physical structure of the system. The first column lists the five issues. The second column presents the physical aspects of the dilemmas, while the third column lists constraints to the solution. The fourth column presents solutions that are theoretically ideal. The fact that they are not feasible is the cause of the five dilemmas, which are listed in the next column. The last column contains some practical solutions, which will be discussed in Section 9.7.

Table 9-1 shows that unbundling – in particular, the loss of control of generation dispatch and investment – is the source of the first four dilemmas. The reason is the lack of a system of efficient incentives for the coordination of generation with the network. The dilemmas are a symptom of the hybrid nature of electricity markets. They result from tension between the goal of creating a competitive generation market and the need to coordinate generation with the network. The inevitable slowness of network development is a significant obstacle to coordinating network development with the generation market. The last dilemma, concerning the need for network capacity to facilitate the market, differs from the rest in that it is not an issue of cost minimization but of the competitiveness of the generation market.

The network manager is the problem owner of the first three dilemmas. The network manager is also the problem owner of the fourth dilemma, although insufficient network capacity may become an issue to generating companies and consumers as well. Network managers are indifferent, however, to a lack of network capacity with respect to competition. In this case, the problem owners are the generating companies that do not have access to certain consumers due to network restrictions and the consumers who do not have the benefit of a competitive provision of electricity.

9.6.7 Consequences of insufficient coordination

Now some of the possible consequences, in the short term and the long term, of insufficient coordination will be explored. A lack of coordination may cause serious economic inefficiencies and complicate network operation and planning. Conceivably, transmission system development may not be able to keep up with changes in the generation market, which could reduce the reliability of service. This section presents a conceptual analysis of the different types of effects and solutions.

The actual costs of insufficient coordination may not become apparent until the liberalized markets have matured. This may take a decade or more, as newly liberalized markets often first need to reduce the excess capacity that was built up by the former monopolistic utilities. Therefore it may take a number of years before new capacity is developed and it may take still another number of years before conclusions can be drawn about the adequacy of the new investment pattern. The costs of insufficient coordination will vary from one electricity system to the next, depending upon the regulatory framework and the physical characteristics of the system. In systems where there are few degrees of freedom for investment in generators, for instance because the main primary energy source is hydropower, there may be fewer coordination issues than in systems

where generators have more freedom.

Short-term consequences

In the short term, sudden unavailability of generation units presents the largest challenge to network managers, both for maintaining power quality and for managing the power flows on the network. The operation of generating units may cease unexpectedly both due to technical causes or for business reasons. As was expressed by the second dilemma, this means that network managers cannot fully depend upon the availability of reactive power management services from generators. They can insure themselves against this risk by installing capacitors for this purpose.

A more difficult risk to hedge against is the possibility of losing a generator that is essential to the reliability of the electricity supply in a certain part of the network. To be sure, if normal reliability standards were applied before its closure, it does not mean that there suddenly is a power shortage. Rather, it may cause difficulty in maintaining these reliability standards, such as the n-1 criterion, which states that normal service must continue if any network link or generation unit becomes unavailable. This means that the probability of a shortage has increased above the norm.

Interim solutions depend upon the situation. If the intended closure was the result of a business decision, an option is to apply a congestion management method that causes the local electricity prices to be higher (congestion pricing), which should provide sufficient incentive to the generator to keep operating. Payments to the generator to remain active are a more direct option. This would be a form of redispatching or counter trading (see Chapter 10), but in an unbundled system the question is whether the network manager has the authority to enter the generation market. Other solutions are to use mobile generators or to implement a local program of interruptible contracts, in which loads are paid for allowing their service to be interrupted during demand peaks. Again the question is whether the network manager is authorized to do these things and, if so, whether he also has an incentive. Thus, it may be concluded that a sudden closure of a generating unit, whether for business reasons or due to technical failure, may threaten both the operating cost of the network and the reliability of service.

Investment strategies

In the long term, the lack of coordination may result in structurally higher system costs than would be economically efficient. These costs may take the form of lower quality of service, given a certain level of system-wide investment, or of higher costs to obtain the quality of service objectives. In a study on the position of network managers in an unbundled, liberalized electricity system, Künneke et al. (2001) identified three possible investment strategies for network managers: network development can be robust, flexible or focused. A robust strategy creates a network that can accommodate a variety of developments in the generation market. It is expensive, as only one of the scenarios will actually develop; investments made in preparation for rival scenarios are at least partly unnecessary. Consequently, by definition a robust strategy leads to over-investment. A flexible strategy is the ideal: an infrastructure design that can adjust quickly to new market developments. Unfortunately, flexible infrastructure technology is more

expensive or simply not available. A significant time lag between market developments and adjustments to the network is inevitable with current technology. The solution chosen most often in the past was to focus the development of the network on one generation scenario. The design of the network was adjusted to the specific locations of the generators. In vertically integrated utilities, this strategy carried little risk; in an unbundled system, the probability that this scenario will be realized has become much smaller. The risk that a focused strategy is based upon the wrong generation scenario could be reduced by providing generators with efficient locational incentives but this is no sinecure.

Now we arrive at the predicament in which network managers may find themselves with respect to the development of their networks: a robust strategy is too expensive, a flexible strategy is not always possible and may also be expensive, and a focused strategy is risky. Consequently, insufficient coordination of generation and the network may leave network managers with a choice between over-investing or taking risks with the future reliability of service. A related risk is that network improvements that appear efficient when they are implemented become obsolete long before their economic life has been reached because their life cycle of several decades makes the development of the networks significantly slower than the development of the generation market. The degree to which these problems occur in practice depends on the specific circumstances of each network, in particular upon the number of degrees of freedom that investors in generation capacity have. If the choice of location of new generators is restricted, be it through economic factors, through geographical limitations or be it through regulatory restrictions, there will be less of a coordination problem.

Coordination between systems

Coordination problems may manifest themselves especially between different connected electricity systems, such as in the western European interconnected grid, rather than within a single system, as was the implicit assumption so far. If the transmission rate is a flat fee that gives the right to transmit electricity to anywhere in the interconnected grid for the same price, Dutch consumers may well choose to purchase their base load electricity from French nuclear plants, their peak load electricity from the hydropower plants in Scandinavia or the Alp countries and their medium load from cheap eastern European plants. However, if natural gas prices decrease, consumers may suddenly turn to the countries near the North Sea to purchase more electricity from natural gas-fired generators.

While this would be good for trade and competition, the question is if these advantages are not more than offset by the increase in network losses and the necessary reinforcements of the international grid. In addition, a strong reduction of generation output in systems with higher generation costs could jeopardize the reliability and quality of service there. Without sufficient generation resources, local network expansion would be necessary to maintain the reliability of service: as generators and network capacity are substitutable for meeting local demand, a loss of generation resources would require an increase in network resources in order to maintain the same level of reliability. There are limits, however, to the extent to which electricity systems can rely upon imports without

becoming physically instable. Services like fault management would also become more difficult with reduced local generation. Currently, trade is limited by congestion on many of the interconnectors between national systems, and in some cases also within systems. However, the EC proposes to require significant expansion of interconnectors (EC, 2001b). It is questionable whether overall system efficiency is served by this proposal.

Distributive aspects

While the networks appear to bear the brunt of the costs of a lack of coordination, there may be costs to generators as well. An example is that inadequate network capacity may limit opportunities for new construction in areas where that otherwise would be attractive. Absent efficient incentives to network managers, the latter may not respond to generator demand for more capacity in specific areas. Another possible effect upon the generation market is caused by the cross-subsidies which inevitably exist if tariffs are not value-reflective. This may result in the favoring of certain types of generation over others, for instance through transmission tariffs that are more favorable to large units, which connect to the high tension grid, than to generators connected to the distribution networks.

Scale of the issue

This chapter merely provides an inventory of possible coordination issues. While some aspects of the issue are evident, such as the issue of congestion in networks with fixed transmission tariffs, the relevance of other issues still needs to be determined. To establish the existence and impact of the possible coordination issues that have been identified, empirical evidence needs to be gathered.

The causes of congestion can be separated into intrinsic and artificial price differences, the latter being caused by differences in taxes, subsidies or cross-subsidization. A useful exercise would be to calculate the price differences among interconnected systems in the absence of distorting taxes, subsidies, differences in the calculation of transmission tariffs, *et cetera*. The resulting flows, which can be considered the product of genuine price differences, should form the basis for interconnector capacity expansion decisions.

To determine the impact of network fees that are not cost-reflective upon the long-term development of the system, a case study should be made of the locational decisions for new generation capacity. The fundamental question to be answered is whether generating companies that do not pay the full network costs of their investment decisions would select different locations if they did have to pay these costs. A similar study should be made of large loads, to answer the question of whether they would provide more generation capacity themselves or locate elsewhere if the network tariffs would have been cost-reflective. If the answer to these questions is positive, and the deviations from the optimum are significant, this is a reason to adjust the regulations and/or incentives.

Finally, modeling of oligopolistic behavior could shed light upon the benefits of network capacity expansion for the sake of improved competition. A method has been established by Borenstein et al. (2000).

Conclusion

The general image that emerges from this analysis is that restructuring shifts costs and risks to the networks for the sake of fostering competition in the generation market, while the control options of network managers are reduced. On the other hand may network managers not always have an incentive to respond to the needs of generators or to consider the social benefits of facilitating competition.

The degree to which the lack of appropriate incentives affects the development of the system depends upon the circumstances of each network. However, it may be concluded that the benefits of competition in the generation market are at least partly offset by a loss of economics of coordination. The question is how to minimize the effects of insufficient coordination. In some cases, a trade-off appears inevitable. In other cases, there may be solutions that approximate the theoretical ideal. In the next section some policy options are explored.

9.7 Policy options

9.7.1 Objectives

In Section 9.3 the general policy goals with respect to the relations between the generation sector and the networks were described. The choice for an unbundled system means that the ideal solution, perfect financial incentives, is not available. Value-reflective financial incentives were shown to be infeasible in decentralized electricity systems due to the existence of network externalities, which was the cause for a number of the dilemmas. Absent theoretically elegant solutions, practically feasible solutions to the dilemmas will now be investigated.

The objective is to create a system of rules and incentives that ensures that the generation sector and the network remain reasonably coordinated. Due to the long lead time for network improvements and the long life cycle of network components, it is unlikely that even a perfectly structured system will be in a perfectly efficient state very often. Therefore the use of pragmatic proxy solutions may not necessarily constitute a great loss of efficiency. The objective is to avoid large inefficiencies – high costs, service disruptions – rather than to achieve a precise optimum. Specifically, the goals are:

1. to be able to influence which generators are active, for the purpose of matching the load flow to the network capacity,
2. to influence the location of new generators, for the purpose of minimization of total investment and the reliability of service, and
3. to ensure the provision of ancillary services, such as reactive power management, for the operational quality of the network.

9.7.2 Instruments

A number of tools are available to achieve these objectives. The analysis of the physical and economic relations between the generation market and the networks in Chapter 2

provides a number of instruments with which the coordination of generation and the network may be improved:

- transmission tariffs,
- network access charges,
- congestion management methods,
- payments to generators (for ancillary services provided), and
- permits, for instance for the construction of generation units or for feeding electricity into the public network.

This section will make a first assessment of the merits of these instruments and their applicability.

Transmission tariffs

Transmission tariffs are the charges that the transmission operator levies for transporting units of electricity from a generation to a load. While the charges are related to electricity sales between two specific points in the network, in the EU, these charges are not allowed to be distance-related (Regulation (EC) No 1228/2003). (In a system of locational marginal pricing, they are not distance-related either but vary with the state of the system.)

The cause of the first, third and fourth dilemmas is the fact that transmission tariffs and network access charges that are fixed *ex ante* are not value-reflective. Absent value-reflective network tariffs, cross-subsidies between network users are inevitable. Given the choice for fixed tariffs, the options for providing incentives with transmission tariffs appear limited to variations in the way the costs of the different voltage levels are allocated and to certain geographic incentives.

The allocation of the costs of the different network voltage levels remains a variable, even if tariffs are not related to distance. Aalbers et al. (1999) outline options for allocating the costs of the different voltage levels among network users. One option is to let the users of the lower voltage levels contribute to the costs of the higher voltage levels, relative to their share of total electricity consumption. In this system, the cost of the higher voltage levels is distributed among all consumers. The rationale is that consumers connected to the lower voltage networks also use the transmission network. This assumes that all electricity production takes place at the highest voltage level, or at least that it is used for the transmission of electricity from the generators to the consumers. In systems with distributed generation, this assumption is not necessarily true. Electricity produced by small generators and fed into the distribution grid typically is used by consumers within that network. Therefore generators close to consumers reduce network losses and the need for network capacity. However, these benefits are not reflected in the price of their electricity, as these plants still need to contribute to the costs of the transmission network, in addition to the tariffs for the use of the distribution network to which they are connected.

An alternative, described by Aalbers et al. (1999), is to allow consumers who purchase from generators connected to their own distribution network to pay a tariff that is only based upon the local network costs. However, in systems with a large share of distributed

generation, this may result in insufficient funding for the transmission network. This may raise equity issues, as the transmission network provides more services than the transport of energy alone. Most importantly, the connection of local networks to each other greatly enhances the stability of the electricity system. It also improves the quality of the electricity, as the frequency is more stable in larger networks. These benefits are public goods and therefore difficult to capture in transmission tariffs.

Geographic incentives may be provided through variations in the connection charges or through the use of congestion management methods. These options will be discussed below. Another option is to apply the flat transmission tariffs only within specific zones, while charging a fee for crossing from one zone to the next. This reduces inter-zonal trade to those transactions with a surplus higher than the cross-border tariff. This is the current situation in the EU, with the member countries constituting the zones (EC, 2002b). The cross-border charge is levied to raise revenues to pay transit countries but it also sends a signal – albeit a primitive one – that electricity from a remote source costs more. Both the choice of the zonal borders and the level of the fee are somewhat arbitrary, so this type of levy can not be economically efficient. However, it may still be of use to limit highly inefficient behavior such as purchasing electricity from a remote source for a price advantage that is lower than the associated transmission costs.

Network access charges

Network access charges are the price that generating companies or consumers pay for being connected to the network. These charges are not related to the quantity of electricity produced or consumed but they typically vary with the capacity of the connection. Network access charges provide more options for providing efficient investment incentives. Network access charges can be varied among generators and/or consumers by location, capacity and voltage level. This would not increase transaction costs, while the increase in the complexity of the market remains limited. Geographical variations in network access charges are being used in systems with a substantial, one-dimensional cost differential, such as England and Sweden. Both these systems have a shortage of generation capacity in the more populous south and excess capacity in the north, partly due to the location of hydropower plants. By increasing network access charges in the north and decreasing them in the south, they stimulate new generators to locate closer to demand, thus reducing network losses as well as the demand for capacity on the north-south transmission links. However, when access charges are used to influence electricity flows between two systems, they may encounter opposition from consumers if the effect is to make electricity in the exporting system more expensive.

It may also be difficult to fine-tune the incentives created by network access charges. The cases of England and Sweden are rather straightforward, with a simple cost differential between geographically remote areas. However, the deep connection costs of a new large load may vary considerably over relatively short distances. For instance, if the electricity grid on the west side of a city is near capacity, while there is ample capacity to its east (for instance due to the proximity of the transmission grid), the costs to the network of a new large load on the west side may be much higher than on the east side. Vice versa, generators who locate in the congested area on the west side lower the demand for

transmission capacity and thereby reduce system cost. Converting these cost differences into locational incentives would require an intricate system of access charges that would vary over short distances.

Another issue with variable access charges is the question how they should be varied over time. Given the irreversibility of investment in generation capacity, should generators face variable access charges, or should they be determined for the life time of a generation unit? Variable charges would provide a dynamically more efficient incentive but their unpredictability may reduce their effectiveness in influencing investment decisions.

Stoft (1999) offers a critique of the use of locational access charges that apply to all generating companies. He argues that the charges will not cause existing generators to move but they may cause them to leave the market or, in the case of positive incentives, provide windfall profits. Stoft (1999) proposes to target incentives only to new capacity. He argues that the transmission operator should offer positive incentives to new generators that reduce the need for transmission capacity expansion, rather than impose charges upon generators that contribute to congestion. By letting generators bid competitively, the incentives can be limited to an efficient level. The fact that the transmission operator pays the generators provides him with an efficient incentive to balance network capacity upgrades with investment in generation.

Congestion management methods

A natural way of providing locational incentives is through congestion pricing. The presence of congestion means that the combined result of all market transactions would cause electricity flows in excess of the available capacity, so some form of intervention in the output pattern of the involved generators is required. Congestion pricing methods divide the market into different price areas, separated by the congested links. 'Downstream' of the congestion the prices are highest, so there is an incentive for generators to increase their output and to invest in new units. As the price difference between the upstream and the downstream markets changes, so does the market value of the congested link.

Congestion pricing is an effective, market-oriented way to allocate scarce transmission capacity. As a corollary, it provides efficient locational incentives, at least in theory. A limitation is that congestion pricing methods can only be applied to structural cases of congestion, as they require specific institutional arrangements. A second limitation is that they divide the market into zones with different prices, which may present an obstacle to trade. The specific effects in terms of efficiency, trade barriers, transaction costs, et cetera are discussed in Chapter 10, which is dedicated to congestion management methods.

Permits

If economic incentives are not an option, it may be necessary to intervene more directly in the market if the quality of service or the overall system efficiency are threatened. Possibilities to intervene in the generation market are limited in a liberalized system but one option is to use construction and operating licenses for generation units to impose

conditions upon the generators, for instance to guide the location of new generators.[60] However, involvement of the network companies in the permitting process could compromise their independence, as they now have a way to intervene in the generation market. Only in completely unbundled systems could this be an option; otherwise, the power to deny generator investments would provide too strong a competitive advantage. As the interests of generating companies and network companies may be opposed, there is need for a counter balance to ensure that not only the network companies' interests are furthered. A regulatory review process, based upon benefit-cost analysis, could balance the interests. This, however, is a step back towards the central-planning paradigm.

Permit requirements can also be used to require a reasonable price for ancillary services, if the generators have a local monopoly over these. The application of permits, however, is limited in scope, as they can only prohibit activities or impose conditions, whereas incentives can also have a positive force.

Locational marginal pricing

Starting point of the analysis in this chapter was the choice for transmission tariffs that are fixed *ex ante*. This precludes the use of locational incentives that vary in real time such as locational marginal pricing. However, the complications that arise from this choice merit a re-evaluation of the possibilities to introduce a form of real-time locational incentives. Perhaps it is possible to implement locational marginal pricing within each electricity system and to provide efficient congestion management methods between the systems. Alternatively, zonal pricing may be a first step, with each electricity system constituting only one or a few zones. These options are outside the scope of this chapter but the existence of alternatives to fixed transmission tariffs should not be forgotten.

9.7.3 The limits of incentive regulation

The approach taken in this section was to search opportunities for economic incentives where possible, in the spirit of liberalization, even if theoretically efficient incentives are not feasible. The analysis reveals some weaknesses of this approach, however. A limitation is that incentives function on a different time scale than the development of the sector. The capital-intensiveness of both generation and network means that their capacity does do not respond to short-term fluctuations of price signals. Rather, investment decisions should (ideally) be based upon forecasts of the value of these incentives over the economic life span of the investment. In practice, investment may be backward-looking and based upon the recent experience.[61] The result may be that the incentives are less effective than one would expect. In combination with risk-aversion among the investors, this effect may be lop-sided: it may be possible to deter investment in certain cases through high charges, while it may be much more difficult to stimulate investment in other cases. Thus, there is a risk that guiding investment through incentives leads to a permanent lag between network development and market demand.

[60] Network companies may do this anyway, unofficially, by not cooperating with generator locational decisions which they do not agree to.

[61] This argument is analogous to the argument, presented in Chapter 5, that a competitive energy-only market may not lead to an optimal volume of generation capacity.

Even from a static equilibrium perspective, the use of incentives contains an inherent risk of sub-optimal development because it is difficult to create perfect incentives, at least in a system with separate transmission tariffs. When incentives are not economically efficient, it is to be expected that the system equilibrium will not be socially optimal. This risk is magnified when the incentives are not theoretically efficient at all but consist of *ad hoc* price signals. Since such signals are not intrinsically efficient, they need to be recalibrated periodically. This, in turn, creates uncertainty about the future value of the incentives, which may undermine their effectiveness. Here we come upon a fundamental paradox regarding the liberalization of the electricity sector: for the sake of improving economic efficiency (introducing competition), liberalization involves a shift from performance regulation of the electricity supply industry to process regulation (establishing the framework and the rules for the electricity market). However, the difficulty with establishing economically efficient process parameters threatens the goal of economic efficiency. Section 11.5 reflects upon this issue.

9.7.4 Conclusion

All but the last of the five dilemmas that were presented in Section 9.6 may be dealt with through variations in the financial structure of the transmission tariffs, connection charges and congestion management methods, and through the rules set out in connection or operating licenses. These instruments can be used to provide operational and investment incentives to generators, which may approximate the effects of efficient incentives. Licenses may also be used to prevent the abuse of monopoly power over ancillary services. The threat of large economic efficiencies due to a lack of coordination may thus be reduced.

The incentives given through these measures do not reflect real costs but are proxy incentives which are intended to compensate for the lack of efficient transmission tariffs. The fact that they are not inherently efficient means that they need to be recalibrated periodically. It also means that they are only an approximation of the ideal incentives, and that they may create new external effects. They clearly are a second-best option, a solution which should be used sparingly only to compensate for the absence of a more elegant solution.

Due to the short history of liberalized electricity systems, insufficient insight exists with respect to the scale of the issues that were described here. While all of them are likely to occur to some degree, more research is needed to estimate the cost of the associated inefficiencies. Similarly, the effectiveness of the proposed solutions has not been tested: this chapter has merely presented a number of options.

Liberalization of the electricity sector appears to have reached the limits of guiding actor behavior through competition and incentives. The problem is two-fold: theoretically efficient incentives should create a socially optimal system equilibrium but it is unlikely that the system ever is in a state of equilibrium. Secondly, it appears impossible even in theory to create efficient incentives in a system with fixed transmission tariffs. This poses a fundamental paradox of liberalization: an important goal is to enhance economic efficiency through the introduction of competition and other economic incentives but the

inevitable imperfection of these incentives reduces economic efficiency. To reduce the risk of all too great deviations from the socially optimal system structure, in some cases it may be necessary to revert to more planning-type solutions, such as benefit-cost analyses for network expansion.

9.8 Paradigm shift

In the process of liberalization, European countries have chosen to separate the generation market from network operation. In order to stimulate competition, market parties are free to engage in bilateral contracts or to trade in organized power exchanges. To maximize the transparency of the market, the tariffs for the use of the network are made simple and equal among large categories of users.

It may be argued that the main existing alternative, locational marginal pricing, is not feasible in the current European system because it would require a high degree of institutional harmonization among the different national electricity systems. While the choice against locational marginal pricing in Europe is legitimate and perhaps, at least in the near term, inevitable, current European electricity policy limits the options for coordinating the development generation and the networks efficiently. In the first years after liberalization, this omission may not manifest itself strongly because the EU countries started with a large excess of generation capacity, as a result of which there is little need for new construction of generation units. Initially, the only visible effects will be in the operation of the network, where consumers may make inefficient purchasing decisions, which would lead to higher energy losses and an increased probability of congestion.

Once the transition phase has passed and the excess capacity has been eliminated, inefficient locational decisions for new generation capacity may create significant costs for the networks. Absent efficient locational incentives, generator investment and closure decisions may reduce the reliability of service locally and impose unnecessary costs upon the networks. Whereas the need to provide locational incentives is recognized (Regulation (EC) No 1228/2003), there also is a strong push to increase interconnector capacity to accommodate these flows (EC, 2001b). This opens the prospect of large electricity flows across the continent which may change rapidly when market conditions change. A substantial expansion of network capacity would be required to accommodate these large flows. The objective of a minimum interconnector capacity of 10% appears arbitrarily chosen (EC, 2001b; European Council, 2002).

It is questionable whether the gains from competition warrant these extra costs. At least the trade-off between the costs of facilitating competition in the generation market and the benefits of competition should be made explicitly. This requires a paradigm shift in European policy with respect to network regulation. Policy has focused on third party access (TPA) issues, with the facilitation of competition as a main goal (EC, 2001a). The ideal of the current policy sometimes is described as a continent-wide ‘copper plate’, which allows electricity to be injected and withdrawn at any location without capacity restrictions. This does not only require reform with respect to network access but also

significant investments in the capacity of the network. The danger of this policy is that it confuses means with ends: competition is a means to achieving economic efficiency, not an end in itself. A policy of maximizing competition carries a risk of losses to economic efficiency elsewhere in the value chain. The cost of network expansion may not always be balanced with sufficient benefits in terms of increased competition or improved reliability. The result would be a less-than-maximal gain in economic efficiency, or perhaps even a net loss. In this respect the metaphor of the copper plate is appropriate: a copper plate the size of the continent would not only be good conveyor of electric energy, it also would be expensive.

To avoid large economic inefficiencies in the long run, European policy should depart from the goal of a continental copper plate and recognize that electricity is somewhat of a local product.[62] Transmission causes energy losses while the local presence of generation capacity facilitates many aspects of system management. Congested interconnectors, for instance, therefore are not just a barrier to trade; congestion may also signal efficient investment incentives. Rather than striving for a European network that facilitates cross-continental competition, policy should return to the original goal of system-wide economic efficiency. Proxy incentives may reduce the worst inefficiencies without changing the basic structure of the system. The negative impacts upon competition would be limited, as modifications of existing charges could be used.

The absence of a theoretically sound incentive system may not constitute a large disadvantage in practice. The difference between the time constants of the electricity market, the generation stock and the networks – the high speed at which market prices change, versus the life cycle of generators, versus the even longer life cycle of networks – means that the effectiveness of economic incentives to guide the long-term development of generation and the networks may be limited, even if they are perfect in theory. Perhaps additional regulation is inevitable to secure against the worst economic losses from insufficient coordination. This additional regulation may take the form of permit requirements for generators, but also of a requirement of government approval for large network expansions, for instance based upon cost- benefit analyses. Because a true optimum will probably never be reached, the loss to efficiency due to the pragmatic approach advocated in the previous section may be limited, compared to a system with theoretically perfect incentives.

9.9 Conclusions

This chapter reviewed the relationship between generation and network. One of the principal goals of liberalization is to improve the economic efficiency of the sector; the principal means to this end is to create competition wherever possible in the sector. As competition can only be introduced in part of the electricity sector, there is a risk of insufficient coordination between the regulated monopoly functions and the competitive functions. The question this chapter addressed is how to coordinate the development of

[62] Since the original version of this chapter was published (De Vries, 2001), the need for locational signals has been recognized (EC, 2003).

generation and the network. In addressing this question, some difficult choices need to be made.

Unbundling and system optimization

The cause of the dilemmas is the need to unbundle the networks from the generation market for the sake of fair competition, while the two are physically closely related. As a result, unbundling comes at the cost of reduced opportunities for coordination. Generators and the network are related with respect to:

- load flow,
- network voltage (reactive power management),
- long-term development (investment, withdrawal of capacity), and
- competition in the generation market (because this is influenced by network capacity).

The paradox of fixed transmission tariffs

In liberalized systems, the preferred coordination method is through a system of economic incentives that mitigate the effects of externalities. The only theoretically sound system that may provide adequate incentives is locational marginal pricing. However, the EU favors fixed transmission tariffs, which are simpler and more transparent than locational marginal pricing, and easier to implement in an interconnected system in which the member systems differ in many respects. Unfortunately, the existence of network externalities makes it impossible to create theoretically efficient transmission tariffs that are known *ex ante*.

Nevertheless, the need for coordination mechanisms between the generation market and the network remains. Absent economically efficient incentives, these mechanisms necessarily will be a mixture of *ad hoc* measures. As these proxy incentives lack a theoretical foundation, they will need to be evaluated and recalibrated periodically. Such measures complicate the market, reduce transparency and create a certain degree of regulatory uncertainty, undermining the desired transparency of the fixed transmission tariffs. Thus, the paradoxical situation develops of a complicated, intransparent combination of *ad hoc* fixes to mitigate the negative side-effects of fixed transmission tariffs, which were established to provide simplicity and clarity.

The ultimate question is whether the benefits of competition in the generation market merit the extra costs that are caused by unbundling and other measures to facilitate competition. This question, however, may never be answered, as the full costs and benefits of liberalization will only become apparent after many years, and by then it may no longer be clear how the former integrated system would have performed.

Dilemmas

In addition to the problem of network externalities, there are some practical obstacles to finding a balance between the costs and benefits of facilitating competition:

- the generation market operates in a shorter time frame than the networks do,
- the generation market may develop temporary or intermittent monopolies, and

- it may be difficult to estimate the benefits of increased competition over the life cycle of network investments that were made to stimulate competition.

As a result, five dilemmas were identified:

> *Dilemma 9.1: Value-reflective network charges fluctuate unpredictably and are therefore unacceptable to network users, unless they are integrated in the energy market. By definition, this is not possible in a decentralized system, but network charges that are not value-reflective create significant externalities*
>
> *Dilemma 9.2: For the provision of reactive power, network managers have the choice of relying on local generators, which is risky, or investing in capacitors, which is more expensive.*
>
> *Dilemma 9.3: Unbundling prohibits integrated planning of investment in generation and network capacity by a single firm, while the varying nature of network costs hamper the creation of efficient incentives through fixed transmission tariffs and network access charges.*
>
> *Dilemma 9.4: Due to the long lead time for network investment and the long life cycle of networks, network investment in anticipation of market developments is risky and therefore involves higher average costs; however, reacting to changes in market demand means being substantially too late, which also creates high social costs.*
>
> *Dilemma 9.5: To what degree should network capacity be increased to enhance competition when market conditions, and hence the benefits of such capacity improvements, may change on a much shorter time scale than these network improvements take place?*

Coordination mechanisms

The fifth dilemma can, in principle, be regarded as a trade-off between the costs of network improvements and the social benefits of increased competition, where the dilemma stems from the fact that the latter are difficult to project over the life cycle of the network improvements. The first four dilemmas are a matter of coordination, for which several practical mechanisms have been identified. Examples are separate congestion management methods for interconnectors, variations in connection tariffs to provide locational incentives to generators and permit requirements for generators. Such measures compromise the goal of simplicity and transparency, but in practice their effects may be less different from the effects of perfect incentives than one might assume. Incentives are only one of a number of factors that contribute to investment decisions; other factors may still lead to sub-optimal investment decisions, such as insufficient information about future market conditions, regulatory uncertainty and the long lead time for both generation and network expansion. As a result, even a system with theoretically perfect incentives may need additional regulation.

9.10 Recommendations

9.10.1 Policy recommendations

To avoid economic inefficiencies in the long run, European policy should recognize that electricity is somewhat of a local product. There is a need for a paradigm shift in the policy for the electricity market from a focus upon competition, which after all is only the means, to a focus upon overall economic efficiency, which is the end. Supplementary measures will be needed to adjust the long-term development of the generation market to the physical capabilities and the cost structure of the network. Two strategies are possible. The first is to refine the current system of fixed transmission tariffs with *ad hoc* incentives and regulations to minimize the negative externalities. This will add complexity, but does not require a fundamental policy shift. The alternative is to explore options for locational marginal pricing, for instance by starting with zonal pricing.

9.10.2 Research recommendations

This chapter presented an analytic framework for understanding the coordination issues. The next step is to quantify the issues by gathering empirical evidence. This can be supported with combined technical and economic modeling. The effects are certain to differ substantially from one system to another, for instance, hydropower plants do not come and go with the same speed as small gas-fired units. The following kinds of information should be collected:

- the extra cost to network managers of providing reactive power services themselves,
- the vulnerability of network management to the sudden closure of power plants,
- the impact of artificial price differences upon inter-system flows, and
- the impact of non-cost-reflective network fees upon the locational decisions of generators.

The different options for coordinating the generation market with the networks should be evaluated. The options should be assessed in combination with each other, with the goal of developing a feasible, effective and efficient package. Finally, the advantages and disadvantages of nodal pricing should be compared to a system with fixed transmission tariffs, supplemented with the necessary additional incentives and regulations. This means that the choice for a decentralized design of the electricity system should be reconsidered.

10 Congestion management

Electricity networks with fixed transmission tariffs are prone to congestion. As physical overloading of the network must be avoided, a number of congestion management methods have been proposed. Ideally, these methods do not only provide a means to adjust generator output to the physical capabilities of the network but also provide efficient short and long-term incentives to both generators and network managers. In this chapter a simple economic model is used to compare four congestion management methods: explicit and implicit auctions, market splitting and redispatching.

10.1 Introduction

This chapter will review one category of solutions to the issue of coordination that was raised in Chapter 9. When a choice has been made for transmission tariffs that are fixed *ex ante*, as generally is the case in the decentralized electricity systems of Europe, congestion management methods provide a way for network managers to intervene, directly or indirectly, in the dispatch of generators. Fixed transmission tariffs represent an average cost (averaged over time and geographically), rather than the marginal cost of use of the system. This means that they do not provide efficient operational signals to market parties, as a result of which the network may not be able to accommodate all scheduled transactions. To allocate the available transmission capacity a congestion management method is needed. Congestion management methods therefore are *ad hoc* remedies for the lack of incentives provided by fixed transmission tariffs. This model is fundamentally different from an integrated system with locational marginal pricing, in which electricity is traded through a mandatory pool which includes congestion management in the dispatch of generation. Locational marginal pricing currently is the only system in which efficient incentives for the use of the network are built into the network tariff system but for reasons described in the introduction to Chapter 9 it does not appear an option for the management of congestion in Europe in the short term.

The analysis in this chapter place much analysis on the economic efficiency of congestion management methods, as improving the economic performance of the system is an important goal of liberalization (EC, 2001c). Congestion pricing methods meet this objective best, as they make the value of the network explicit. This chapter describes

three congestion pricing methods: explicit auctions, implicit auctions and market splitting. The fourth congestion management method, called redispatching, does not provide efficient incentives to market players but is included because it provides an interesting comparison.[63] It is the default congestion management method. It is more flexible than the congestion pricing methods, for which reason it often is used together with them to fine-tune the interconnector capacity allocation in real-time. In addition, a theoretically interesting aspect is that redispatching provides more efficient incentives to the network manager than congestion pricing methods. Redispatching can be considered a 'corrective' congestion management method, as it does not influence the market, but leaves it to the involved TSOs to correct the situation.

Three short-term economic aspects of each method are discussed. First, short-term economic efficiency. Second, how costs and benefits of each method are distributed among the producers, consumers, network operators and market operators. And finally, attention is given to more practical economic aspects, such as transactions costs and suitability for application in complex (meshed) network. Following the description of the individual methods, the latter part of the chapter is devoted to a comparison and general analysis, among others, of the incentives they provide for generation and network investment.

This chapter shows that all the investigated congestion management methods can, in theory, achieve economic efficiency in the short term. This means that they result in the most efficient dispatch of generators, given demand and transmission constraints. Their differences lie in the question of whether they are efficient in practice and the economic incentives they provide for long-term development. These issues will receive special attention in the last paragraphs of this chapter.

This chapter is structured as follows. The next section presents the model that is used for the analysis of the congestion management methods. The methods themselves are described in Section 10.3. A reflection upon the economic assumptions that underlie the analysis is presented in Sections 10.4. Section 10.5 discusses the impact of a particularly strong assumption regarding the physical structure of the network and a more refined approach which currently is being developed. Section 10.6 provides a comparison of the different congestion management methods with respect to their short-term economic efficiency, welfare effects and long-term incentives.

10.2 Analytic framework

For the analysis of congestion management methods, this section introduces a general model to describe the simple case of electricity export from country A to country B over a single interconnector. First some basic assumptions are introduced, then a model is presented with which the congestion management methods are described, and finally some reference cases are described to frame the analysis.

[63] Knops et al. (2001) provide a general introduction and analysis of all four of these congestion management methods.

10.2.1 Assumptions

In the following sections, for each congestion management method a theoretical, economic analysis will be made, which shows the potential for reaching economic efficiency and the potential economic effects for generators, consumers and transmission system operators (TSOs). The analysis is based upon the ideal case of perfect competition, and therefore does not enter into practical problems such as strategic behavior by market parties and imperfect information. Assuming perfect competition means, among others, assuming that

- all players have perfect information,
- there are many market players, so no individual player can influence the market price,
- market parties can freely enter and exit the market, and
- the product is homogenous.

Only the last assumption holds in electricity markets, and some markets may have enough players. The other assumptions do not hold: there always is a difference in the information to which market players (and government) have access. Free entry and exit does not exist in the power market. On the production side, the high capital requirements obstruct easy entry and exit of the market; on the demand side, nearly all consumers are so dependent upon electricity that they cannot leave the market.

Nevertheless, the model of perfect competition is used here because it is a convenient and widely accepted starting point for an analysis of market systems. This simplified way of looking at congestion management helps to understand the basic structure and the potential of each proposed method. A next step, of which this chapter makes a beginning, is to assess the impact of market imperfections upon each system.

10.2.2 Model

Assume two countries A and B, with electricity supply curves S_A and S_B. These supply curves represent the combined marginal cost of production of all generators in the respective countries. They can be found by combining the marginal production cost curves of all individual generators, when they are ranked from cheap to expensive. Electricity is cheaper in A, which means that S_A is smaller than S_B for output values close to demand.

Figure 10.1 shows the supply and demand curves. The countries' demand functions D_A and D_B can best be interpreted as willingness-to-pay functions. Physical demand is on the X-axis; the demand functions indicate the corresponding prices that consumers are willing to pay. The intersections of the supply and demand curves represent the equilibrium points for each country. On the vertical axis are price (P) and cost (C). The variable on the horizontal axis is Q, which stands for quantity. Output and demand both are a function of Q. Values of Q represent quantities of electricity that are either produced or consumed.

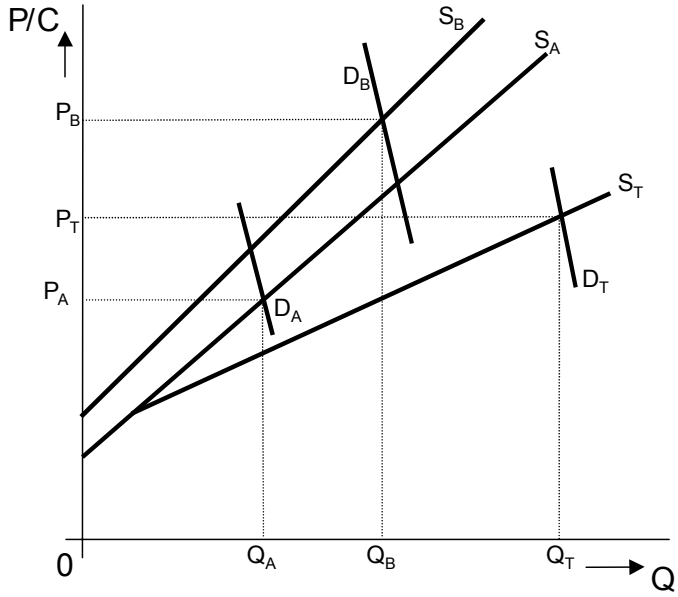


Figure 10.1: Basic model with supply and demand curves for the two countries A and B and for the joint market T

If the markets are joined, a new equilibrium will develop. Producers can sell in both countries, consumers can buy in both countries. For the model, this means that the supply functions of *A* and *B* are added to form a new supply function S_T (*T* for total). The new demand function can be found the same way. The result from joining the markets is that the more expensive generators in *B* are out-competed by the generators in *A*.

The new supply curve is found by adding the curves of S_A and S_B horizontally (along the *X*-axis). For each market price, supply is determined by adding the supply in *A* and the supply in *B* that correspond to that price. The new function S_T coincides with S_A in the beginning. When S_A reaches the price level at which S_B begins, S_T branches off.

Mathematically, the combined supply curve S_T can be obtained the following way. For horizontal adding, the inverse functions need to be used. So:

$$S_T^{inv} = S_A^{inv} + S_B^{inv} \qquad (10.1)$$

Therefore

$$S_T = (S_A^{inv} + S_B^{inv})^{inv} \qquad (10.2)$$

The new demand curve D_T can be found similarly. If demand is assumed to be perfectly inelastic, which it almost is in the short term, total output will be the same in a joined market as in separate markets:

$$Q_T = Q_A + Q_B \tag{10.3}$$

If demand is not considered inelastic, the new output and price are given by the intersection of the new demand and supply curves:

$$D_T(Q) = S_T(Q) \tag{10.4}$$

Solving equation (10.4) renders the total output Q_T. Once this is found, the corresponding market price P_T can easily be found:

$$P_T = D_T(Q_T) = S_T(Q_T) \tag{10.5}$$

10.2.3 Reference Cases

Three reference cases are introduced to provide context for the analysis of congestion management methods. The main focus is on short-term analysis, in which changes to the infrastructure are not possible. The cost of the infrastructure is therefore not included in the analysis. Section 10.6.3 discusses the long-term signals provided by the congestion management methods. The first reference case is the absence of an interconnector. This case should have the highest system cost, as the use of the cheaper generators in *A* is limited. The second reference case is the opposite, namely a situation in which there is an interconnector that is large enough to allow all trades that the market desires. When only the marginal cost of generation is considered, as the only short-term variable, the absence of congestion is the situation with the lowest system cost as all the cheapest generators can run. The third reference case is the case of congestion: there is an interconnector but it has insufficient capacity to accommodate all market transactions. The economically most efficient response to congestion is determined here, without asking the operational questions of how this result can be achieved and what income transfers result. This case provides a reference for the assessment of the economic efficiency of the reviewed congestion management methods.

10.2.4 Reference Case 1: No Interconnector

Figure 10.1 describes the situation in which there is no interconnector. (Ignore the S_T and D_T curves for the moment.) The markets in the countries operate independently from each other. Each market establishes its own price (P_A and P_B) and corresponding output (Q_A and Q_B). Demand in each market (Q_{DA} and Q_{DB}) equals production:

$$Q_{DA} = Q_A \tag{10.6}$$

$$Q_{DB} = Q_B \tag{10.7}$$

Note that demand is indicated by two-letter subscripts, with a D added to denote demand, while production is indicated by one-letter subscripts, only indicating the country.

10.2.5 Reference Case 2: Full interconnection capacity

This case is also represented in Figure 10.1, namely by supply curve S_T and demand curve D_T. Countries *A* and *B* form a single market, with a price P_T and an output Q_T. The output in each country changes as a result of the change in price. The new market price P_T is higher than P_A, as a result of which more of the generators in *A* will run. In *B* the reverse happens: as the price has dropped, a number of generators are priced out of the market. The consumers in *B* now purchase part of their electricity from generators in *A*.

The quantity of demand also changes as the equilibrium changes, at least in theory. In practice, electricity demand is quite price-inelastic, in particular in the short term. This section will first show the analysis assuming demand to be perfectly inelastic. At the end of this section it will be shown what happens if demand is not considered inelastic. In the rest of this chapter demand is assumed to be perfectly price-inelastic.

When demand is considered fully inelastic, this means that consumers in *A* and *B* will always consume the same amount of electricity, regardless of price. Demand will therefore remain the same as production was in the case of no interconnector, as was shown before: Q_A and Q_B. In Figure 10.2, demand is indicated by a vertical line. This means that total demand Q_T can simply be found from:

$$Q_T = Q_A + Q_B \tag{10.8}$$

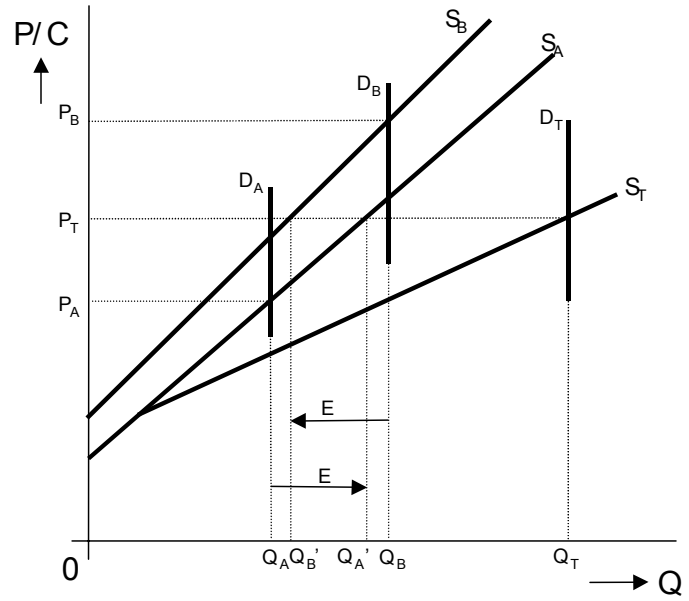


Figure 10.2: Full interconnector capacity, inelastic demand

Figure 10.2 shows the separate supply curves for *A* and *B* and for the combined market. The demand curves are drawn as vertical lines. The new market price P_T can be found

from the joint supply curve S_T at an output of Q_T:

$$P_T = S_T(Q_T) \tag{10.9}$$

The changes in output in A and B can be found as follows. Generation in A increases from Q_A to Q_A'. The prime indicates the situation in a market equilibrium with sufficient interconnector capacity. Q_A' can be found from solving the following equation for Q:

$$S_A(Q) = P_T \tag{10.10}$$

In Figure 10.2, Q_A' can be found as the value of Q that corresponds to a value of P_T for S_A. Similarly, the new output in B can be found from solving

$$S_B(Q) = P_T \tag{10.11}$$

The exported quantity of electricity (E) from A to B can now be determined from the changes in output in A and B. As total demand has not changed (because it is assumed inelastic), the change in output in A equals the change in output in B, and the difference between demand and output equals the export volume.

$$E = Q_A' - Q_A = Q_B - Q_B' \tag{10.12}$$

If demand is considered elastic, the situation becomes slightly more complex. However, the method of determining changes in demand is exactly the same as for changes in supply. As a result of the higher prices for customers in A, demand in A will decrease. The new demand in A, Q_{DA}, is determined by the new market price, and can be found from solving for Q:

$$D_A(Q_{DA}) = P_T \Leftrightarrow Q_{DA} = D_A^{inv}(P_T) \tag{10.13}$$

See Figure 10.3. Note that two indices are used when demand is considered elastic. When demand was inelastic, it was assumed constant and equal to the output level in the case of no interconnector, Q_A. Now the index D indicates demand; values of Q without an index D pertain to generator output levels.

Similarly, demand in B will increase because the consumers in B now pay a lower price. Demand in B can be found from solving for Q:

$$D_B(Q_{DB}) = P_T \Leftrightarrow Q_{DB} = D_B^{inv}(P_T) \tag{10.14}$$

As a result of the elasticity of demand, the decrease in consumption in A is likely to be different from the increase in consumption in B. Therefore the total demand in a combined market may not be the same as the sum of the demand in A and B without an interconnector:

$$Q_{DA} + Q_{DB} \neq Q_A + Q_B \tag{10.15}$$

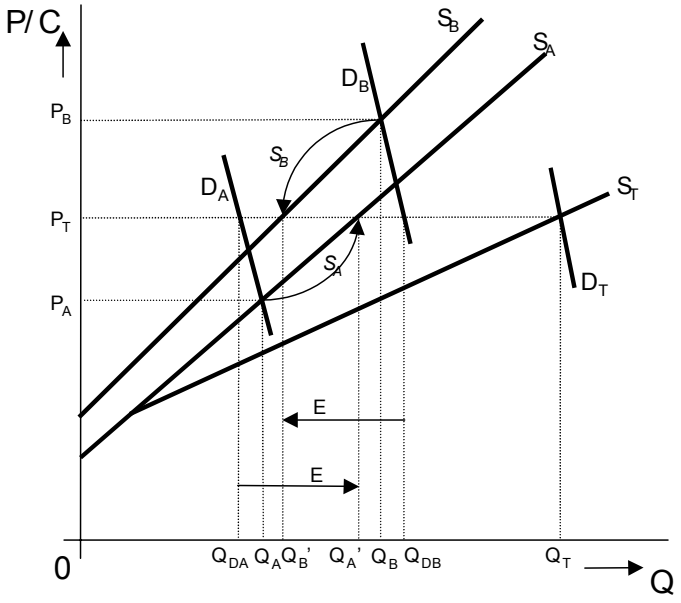


Figure 10.3: Full Interconnection capacity, elastic demand

Total demand Q_T needs to be determined from the intersection of the new supply and demand curves. In other words, Q_T can be found from solving for Q:

$$S_T(Q) = D_T(Q) \tag{10.16}$$

Export E from A to B is equal to the difference between generation and consumption in A, and also to the difference between generation and consumption in B. This is illustrated by the arrows in Figure 10.3.

$$E = Q_A' - Q_{DA} = Q_{DB} - Q_B' \tag{10.17}$$

In the rest of this chapter, demand is assumed perfectly inelastic. In the short term, this is a reasonable assumption which simplifies the analysis.

10.2.6 Reference Case 3: Optimal allocation of scarce capacity

As a third reference case, the theoretically most efficient means of meeting electricity demand when interconnector capacity is less than the market demands shall be considered. The purpose of this optimal allocation case is to establish the lowest possible generation cost. This is used as a comparison for the description of the different congestion management methods: will they also manage to reduce generation cost to the economic minimum? The purpose is not to analyze distributive effects but only economic efficiency.

$$K < E \qquad (10.18)$$

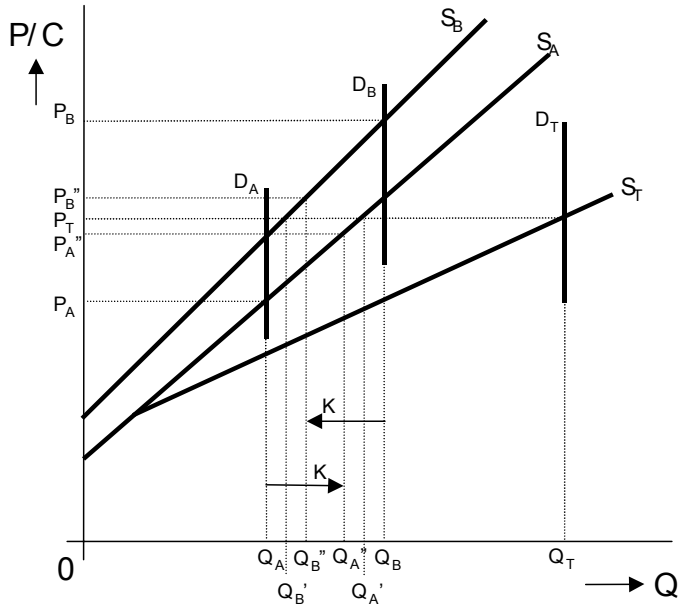


Figure 10.4: Congestion, optimal allocation

As demand is assumed to be fully inelastic, the demand for export is equal to the actual export in the case of no congestion, *E*. The occurrence of congestion means that the capacity of the interconnector *K* is smaller than the demand for export:

As a result, less electricity can be exported from *A* to *B* than the market calls for and equations (10.12) and (10.17) are no longer valid. Generation in *A* will be constrained to a level Q_A'', while generation in *B* will remain at a higher level Q_B''. This situation is represented in Figure 10.4. (The double prime indicates the situation with a congested interconnector. Thus P'' is the price that results in a situation with a congested interconnector.)

Because demand is assumed to be perfectly inelastic, $D_A = Q_A$ and $D_B = Q_B$, also in the presence of the interconnector. The generators in *A* therefore produce an amount Q_A'' that is equal to the demand in *A* plus the capacity of the interconnector; in *B* the generators produce an amount Q_B'' that is equal to local demand minus the interconnector capacity:

$$Q_A'' = Q_A + K \qquad (10.19)$$

$$Q_B'' = Q_B - K \qquad (10.20)$$

10.3 The congestion management methods

10.3.1 Explicit auctioning

This section provides an economic analysis of explicit auctions for interconnector capacity. Explicit auctions currently are a popular method of allocating scarce interconnector capacity in western Europe (EC, 2001c; EC, 2001d; ETSO, 1999). They are currently being used on the border between Germany and Denmark, for the interconnector between France and the U.K., the U.K. – Ireland link, between Italy and Greece, and on the Dutch borders (EC, 2002).

There are several forms of explicit auctions of which *pay-as-bid* auctions and *marginal bid* auctions are the most relevant for electricity markets. These two methods differ in the price that market parties pay for the capacity. Pay-as-bid auctions will be described first, followed by marginal bid auctions. In a pay-as bid auction, each participant who wins capacity in an auction pays the amount he has bid. In a marginal bid auction, the lowest bid that wins capacity (the marginal bid) is the price that all other participants who win capacity also pay. Pay-as-bid auctions are primarily interesting from a theoretical point of view because they generate the largest possible congestion rents. Marginal bid auctions are more often applied in practice because they appear to be more fair and provide better bidding incentives.

First pay-as-bid auctions will be discussed. Figure 10.5, which is an enlargement of the relevant part of Figure 10.4, explains the two types of auctions. The figure shows supply and demand equilibria, with generator output and demand on the horizontal axis and price and cost on the vertical axis. As is described in Section 10.2.5, demand is assumed to be perfectly inelastic, which means that consumer demand in A is equal to Q_A and consumer demand in B is equal to Q_B under all circumstances.

The presence of imports with a volume equal to the interconnector capacity K cause generator output in B to be reduced from Q_B to Q_B''. As Q_B'' becomes the marginal generator, the price in B drops from P_B to P_B''. To supply the exports, generation in A is increased with a volume equal to K from Q_A to Q_A'' so the price increases from P_A to P_A''. As generators in A who export to B want to receive a net income that is at least equal to their marginal cost of production, their willingness to pay at an auction equals the difference between the price they can receive in country B (P_B''), and their marginal cost, which is represented by the curve S_A. The willingness to pay B_g of a specific generator g in A with an output dQ and a marginal cost S_{Ag}, is therefore given by:

$$B_g = P_B''dQ - S_{Ag}dQ = (P_B'' - S_{Ag})dQ \qquad (10.21)$$

The combined marginal cost curve of generators in A who export to B is given by S_A between the points Q_A and Q_A''. Therefore their combined willingness to pay for interconnector capacity is given by the integral of equation (10.21) between these two points. Using equation (10.19) gives:

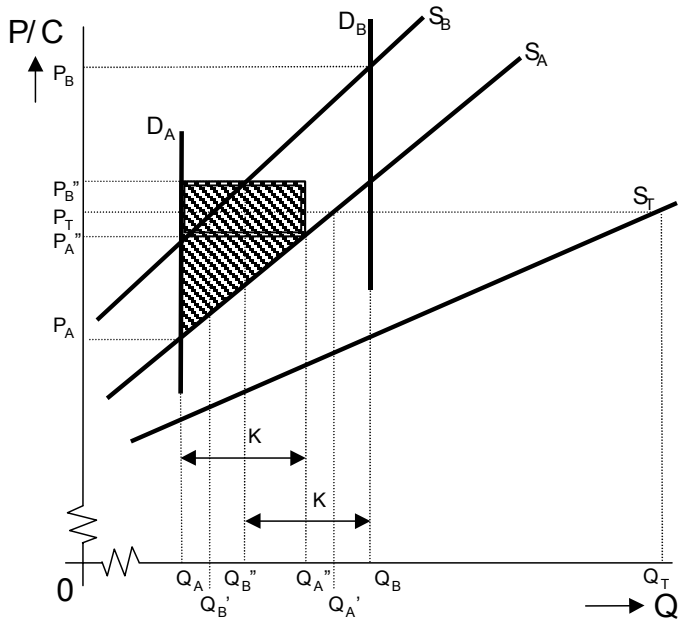


Figure 10.5: Explicit auctioning

$$R = \int_{Q_A}^{Q_A"} (P_B" - S_A) dQ = P_B" K - \int_{Q_A}^{Q_A"} S_A dQ \qquad (10.22)$$

The shaded area in Figure 10.5 represents the combined willingness to pay of all generators who participate in the auction. This amount is in theory equal to the auction revenues in a pay-as-bid auction and represents the theoretical maximum revenue that can be expected from any congestion pricing system. (However, the goal should not be to maximize revenues but to allocate the scarce interconnector capacity efficiently.)

If a pay-as-bid auction works well, the prices paid by bidders ensure that only electricity from the most efficient generators is transported across the interconnector. Therefore it has the potential of achieving economic efficiency in the short term. In practice, bidders who pay more than the marginal bid will see that they could have made a better profit if they had bid less, as long as their bid was above the marginal bid. Bidders will therefore try to estimate the value of the marginal bid, and bid only slightly more themselves. Given the fact that electricity auctions are repeated infinitely, they may become quite good at this game. As a result, the auction revenues will not equal the full willingness to pay of the bidders. In fact, the auction results will resemble less the above case and more the marginal bid auctions.

In a marginal bid auction, the price of the interconnector capacity is set equal to the marginal bid, which is the lowest bid that is awarded transmission capacity. All bidders

who have bid more than the marginal bid receive capacity at the price of the marginal bid. This system reduces the theoretical expected revenues but should result in the same generators obtaining capacity as when a pay-as-bid auction is used. Bidders make a profit equal to the difference between their willingness to pay and the marginal bid. As a result, a bidder's optimal strategy is to bid according to his willingness to pay: if the auction price turns out to be higher, the price is too high for this bidder; if the price is lower, then he is not punished for the fact that he has "overbid". This makes bidding easier and the auction process more transparent than in the case of a pay-as-bid auction. It also improves the incentive to participate in the auction compared to a pay-as-bid auction, so a liquid market is more likely to develop. This in turn may also improve the function of the auction to select the most efficient generators. Therefore a marginal bid auction is also more likely to indicate the market value of the congested interconnector at the time of the auction.

The marginal cost of the last generator to win capacity is equal to P_A'', so he is willing to pay P_B'' - P_A'': the price difference between the two countries in the presence of the interconnector. This is the bid that sets the price for all auction participants in a marginal bid auction. Therefore the auction revenues R will be:

$$R = (Q_A''-Q_A)(P_B''-P_A'') = K(P_B''-P_A'') \qquad (10.23)$$

The revenues equal the price difference, given the presence of an interconnector with capacity K, times the capacity of the interconnector.

In Figure 10.5, the area indicated by the double-lined box represents the revenues from a marginal bid auction. This revenue is always smaller than the revenue from a pay-as-bid auction, unless all bidders have the same marginal cost (in which case S_A is flat between Q_A and Q_A''). Therefore a marginal bid auction should in theory always yield less revenue than a pay-as-bid auction. In practice, the difference may be small or absent due to the incentives for underbidding in a pay-as-bid auction and the incentives for bidding according to willingness to pay in a marginal bid auction, and due to the latter's greater attractiveness to participate. The theoretical congestion rents from a marginal bid auction are the same as from an implicit auction and market splitting, as will be seen in the next sections.

While market parties may pay less in a marginal bid auction, the prices are just high enough to exclude the less efficient generators from access to the interconnector. Therefore this system also has the theoretical potential of being economically efficient in the short term, that is, of achieving efficient generator dispatch. Because this type of auctioning encourages bidders to bid equal to their full willingness to pay, it may actually prove better at selecting the most efficient generators, and may therefore be more efficient in practice then a pay-as-bid auction.

Apart from the choice of bidding system, other important variables in the design of auctions are the time intervals for which capacity is auctioned (days, weeks, months, years) and the firmness of capacity rights. While the above description concerns the theoretical auction results, these other variables will have a significant impact upon the

actual performance in practice.

Both types of explicit auctioning separate energy flows from transmission capacity, which corresponds to the principle of unbundling. An important advantage is that firm transmission access is provided ahead of time. A disadvantage is that this system requires separate transactions for trading electricity and obtaining transmission capacity. The additional transaction increases the complexity of cross-border power trade, and may therefore pose a barrier to trade. When a transaction leads across multiple congested borders, for all of which capacity needs to be obtained in auctions, the complexity increases quickly.

The revenue stream that an auction generates is indicative of the market value of the congested link. It is important to note that this is not an indication of the value of capacity expansion but of the value of the existing capacity. The marginal bid equals, in theory, the marginal value of interconnector capacity. The value of capacity expansion depends upon the additional benefits from trade which it would enable and these depend, in turn, upon the cost curves of the generators that were not able to produce due to the congestion but that would have been in merit otherwise.

10.3.2 Implicit auctioning

In this section an economic analysis of implicit auctioning will be made. An implicit auction is in place on the French-Spanish border, although with somewhat different implementation details than described here. In a system of implicit auctioning, generators in *A* who want to sell electricity in *B* need to bid into an organized spot market in *B*. The market operator increases their bids with a surcharge that is set just so high that the interconnector is not congested anymore. The market operator determines the surcharge as follows. *Ex officio*, he knows both the supply and demand functions in both *A* and *B,* and if he also knows the capacity *K* of the interconnector, he can make an accurate calculation of the market prices in both countries when an amount of electricity equal to the interconnector capacity is exported from *A* to *B*. Looking at Figure 10.6, with interconnector capacity *K* he knows the new output levels Q_A'' and Q_B'', so with the supply and demand functions he can determine P_A'' and P_B''. Because generators in *A* bid into the market in *B,* the market operator must increase their bid prices by such an amount that the market in *B* will only demand for an amount of import that just matches the capacity *K* of the interconnector, which is equal to $Q_A'' - Q_A$. Thus, the marginal generator that is allowed to export is Q_A''. This generator bids a price P_A''. Because the price in *B* will be P_B'', the market operator will set the levy *L* at the difference between the two:

$$L = P_B'' - P_A'' \qquad (10.24)$$

All the generators in *A* that bid into the market in *B* are required to pay this levy. Thus, the revenues from this implicit auctioning are:

$$R = \int_{Q_A}^{Q_A''} (P_B'' - P_A'')dQ = (P_B'' - P_A'')(Q_A'' - Q_A) = K(P_B'' - P_A'') \quad (10.25)$$

These revenues are equal to the revenues from a marginal-bid explicit auction: again the revenues are the interconnector capacity times the price difference. Therefore the same argument holds with regard to economic efficiency: the congestion price is just high enough so only electricity from the most efficient generators is competitive. By excluding the less efficient generators, the congestion management system achieves the goal of creating the most efficient dispatch of generation.

The implicit auctioning process can be considered in a different way that renders exactly the same results. The market operator in *B* accepts bids from *A* in merit order until the interconnector capacity is saturated. The generators in *A* are paid the marginal bid price, P_A''. The market operator sells this electricity in *B* at a price of P_B''. This generates a surplus of the size of the interconnector capacity times the price difference, which was also the congestion rent found in equation *(10.25)*. When considered this way, implicit auctioning closely resembles a one-sided form of market splitting.

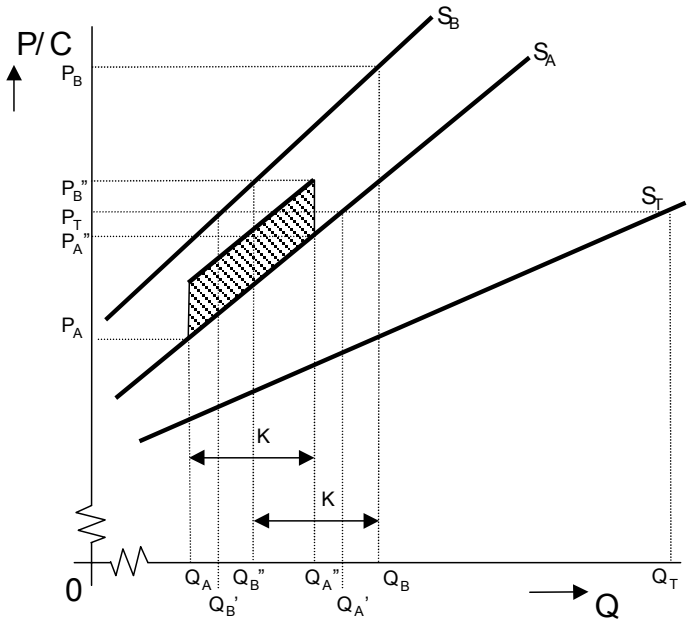


Figure 10.6: Implicit auctioning

Implicit auctioning does not separate energy flows from transmission capacity, which makes the process simpler for market parties. They simply bid into a power exchange and the best bids are honored, until the interconnector is used at full capacity. The revenues that the market operator collects from the fees that he levies should, in theory, be the

same as the revenues from the marginal-bid auctioning. An important difference is that the revenues from implicit auctioning accrue to the market operator, whereas explicit auctions typically are organized by transmission system operators.

The main drawback of implicit auctioning is that it requires an organized electricity market, or at least a market place with a price index, at the downstream side of each congested interconnection. Currently this is not the case everywhere in Europe. A second drawback is that bilateral contracts between the two countries are difficult to incorporate in the system because they hamper the congestion management mechanism.[64]

10.3.3 Market splitting

Market splitting is a third option for congestion management. It is successfully being used in the NordPool area.[65] When market splitting is used to manage congestion, the market is divided by the congested interconnector. There either needs to be an organized market with a separate price on each side of the interconnector, or there need to be two closely co-operating power exchanges. Market parties bid into the organized market on their side of the congestion. In a first step, the two markets are cleared independently. Then the market operator buys electricity from the organized electricity market with the lower price and sells it in the market with the higher price. In doing so, he ensures that the interconnector is used optimally. The result is that the prices in *A* and *B* move closer together (ETSO, 1999).

Because the market operator provides an additional demand in *A*, the market price increases from P_A to P_A''. The reverse happens in the more expensive market: by increasing supply, the market operator lowers the price in *B* from P_B to P_B''. He provides electricity more cheaply than some of the domestic generators can offer. Thus the market operator buys at P_A'' and sells at P_B''.[66] (See Figure 10.7 and Figure 10.8.) The market operator buys just as much electricity as the capacity of the congested line allows: *K*, so his revenues are equal to the price difference times the interconnector capacity:

$$R = K(P_B'' - P_A'') \qquad (10.26)$$

The revenues are equal to those from implicit auctioning and from marginal-bid explicit auctions. The revenues can be earmarked for capacity expansion but they can also be given to the TSO in return for a corresponding reduction of transmission tariffs, as is done in Norway.

[64] The solution that Spain uses is to divide the available import capacity pro rata between bilateral contracts and the spot market, and then arranging separate implicit auctions for both. See www.omel.com.

[65] For a description, see www.nordpool.com/products/elspot/index.html.

[66] The market operator buys electricity at the new market price of the cheaper market, even though he knows all the bids. The rule at organized electricity markets is that all parties receive the same price for the electricity they offer. It would not only seem to be an unjust use of its information for the market operator to only pay the bid price, when the bid price is less than the market price, but would possibly also distort the bidding process in a way similar to a pay-as-bid auction.

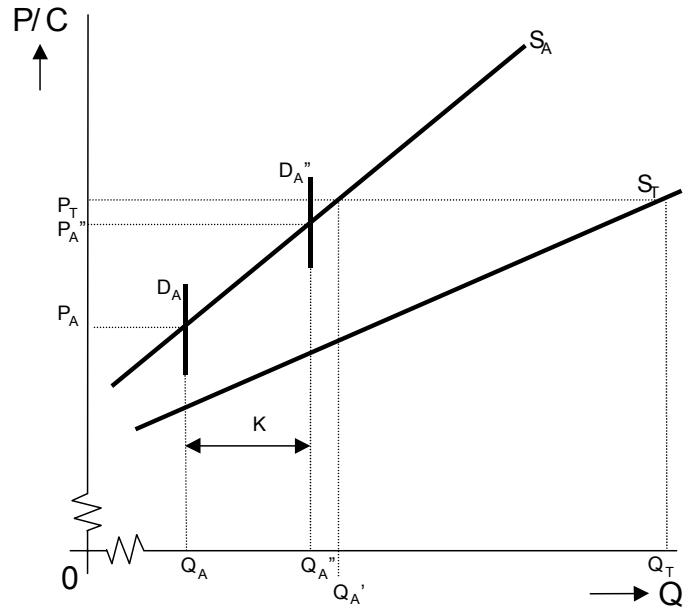


Figure 10.7: Market splitting; situation in A, the lower-priced country

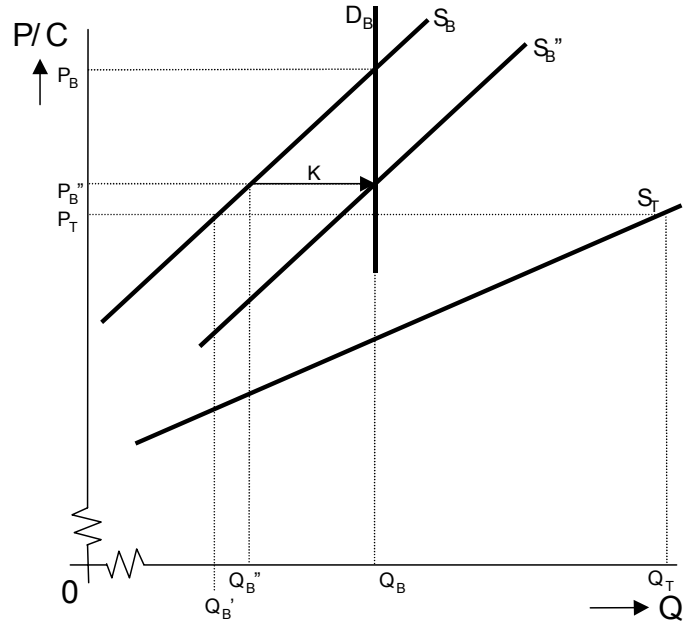


Figure 10.8: Market splitting; the situation in B, the higher-priced country

Graphically, market splitting can be represented as follows. The market operator buys an amount of K electricity in country A. Thus he increases demand by K, so it shifts from D_A to D_A''. See Figure 10.7. By adding an amount of K to the demand curve, the new equilibrium price in A becomes P_A'', at the intersection of D_A'' and S_A.

As the market operator sells the electricity he bought in A to the market in B, he moves the supply curve in B to the right. In Figure 10.8, the new supply curve is labeled S_B''. This causes the price in B to drop from P_B to P_B'', the intersection of the demand D_B and the new supply curve S_B''.

Market splitting leads to an efficient dispatch of generation, simply because the market operator buys electricity from the cheapest generators upstream of the congestion. As a result, generation cost is the same as in the previous cases. Market prices are also the same, so welfare effects are the same too.

This congestion management method has as an important advantage that it is the most convenient of the congestion pricing methods for market parties. They only need to bid into or buy from their own markets. In fact, market parties do not even know whether their bids are used for their local market or for export across the interconnector. Bids are accepted until all demand, including demand from across congested interconnectors, is met. A limitation is that physical bilateral contracts between the two countries are problematic. However, they can be replaced with financial contracts that provide the same benefits to the contract parties. In the Nordic market, the only place where market splitting currently is practiced, there is a strong trend towards financial instead of physical contracts.

Within the EU, market splitting is considered by many as the congestion management method of choice (EC, 2001c). However, the method would need to be adjusted for the highly meshed network of continental Europe. NordPool has a relatively simple structure with few parallel paths between price zones. To adjust it for the European mainland, a solution would need to be found for the existence of parallel flows, similar as is being proposed for explicit auctions (see Section 10.5). A system of market splitting for the European interconnected system, with all trade between the many zones taking place through a centrally operated market, would begin to resemble locational marginal pricing.

10.3.4 Redispatching

In this section redispatching and its variant counter trading will be discussed (ETSO, 1999). Redispatching is a corrective method: it allows the market maximum freedom and leaves it to the TSO to correct any resulting congestion. There are different ways to implement redispatching. Here a system is described in which the market experiences as little as is possible of the existence of congestion. It works as follows. The market trades as if interconnector capacity is unlimited. As a result, a single price develops in A and B: P_T. The market price and corresponding generator output are found as described in the reference case titled Full Interconnector Capacity (Section 10.2.5). This results in a demand D_e for electricity flow across the congested line that exceeds its capacity K.

Contrary to the congestion management methods that have been described until now, the

market is not required to change its transactions as a result of the existence of congestion. Therefore the generators in *A* receive market contracts to produce at level Q_A' and those in *B* at level Q_B' (see Figure 10.9). Clearly, this would lead to a net flow from *A* to *B* in excess of the interconnector capacity. To avoid physical overloading of the interconnector, the TSO intervenes directly in the generating pattern in both countries. He reduces output in the exporting country and increases generation in the importing country up to the point that the net flow across the interconnector matches the available capacity. This process of adjusting generation output by the TSO is called redispatching. In the case of congestion on interconnectors between different systems, the term cross-border coordinated redispatching is used (ETSO, 2001b). The two (or more) involved TSOs need to cooperate closely to make this work.

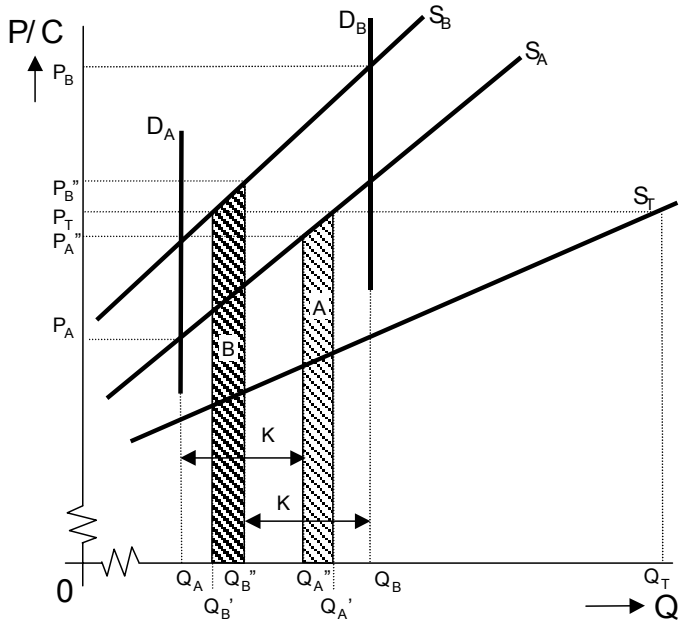


Figure 10.9: Redispatching

Redispatching, as it is defined here, costs the TSOs money. This can be seen as follows. In *B,* the importing country, the TSO needs to dispatch additional generation. He can choose from the generators that have received no market contract to run in the unconstrained situation. Under the conditions of perfect competition, the only reason they would not run would be that their marginal cost is higher than the market price P_T. Of course, the TSO will still try to find the cheapest remaining generators. In Figure 10.9, they are the ones between Q_B' and Q_B''. A generator *g* in country *B* with an output *dQ* receives revenues $T_{B,g}$ from the TSO:

$$T_{B,g} = S_{B,g} dQ \qquad (10.27)$$

All generators in *B* that are dispatched by the TSO receive just their marginal cost of operation. Together they receive:

$$T_{TSO \to B} = \int_{Q_B'}^{Q_B''} S_B(Q) dQ \qquad (10.28)$$

In Figure 10.9 this integral is represented by the shaded area with label *B*.

The opposite happens in *A,* the exporting country. Here production must be reduced from Q_A' to Q_A'' in order to bring the exported volume of electricity within the capacity limit of the interconnector. The decrease in *A* is equal to the increase in *B* (because demand is assumed inelastic). The generators in *A* whose output is reduced still receive the unconstrained market price for their generation contracts (which are still valid) and by not generating they save their variable costs of production. The TSO demands a reimbursement from these generators in *A* equal to their avoided costs, so the redispatching process leaves them financially indifferent. The reimbursement by generators in A to the TSO, $T_{A \to TSO}$, is determined analogously to the payments of the TSO to the generators in *B,* equations (10.27) and (10.28):

$$T_{A \to TSO} = \int_{Q_A''}^{Q_A'} S_A(Q) dQ \qquad (10.29)$$

In Figure 10.9, this is indicated by the shaded area labeled *A*.

As all the generators in *B* that are commissioned to run by the TSO have a marginal cost of production larger than P_T, while the cancelled generators in *A* all were competitive, and thus had marginal costs of less than P_T:

$$T_{A \to TSO} < T_{TSO \to B} \qquad (10.30)$$

From this follows that the reimbursement that the TSO receives from the cancelled generators in *A* will always be less than the cost of the extra generation in *B*. This is logical: had the generators in *B* been cheaper, they would already have run. Therefore redispatching will always cost the TSO. The costs typically are recovered through the general transmission tariffs.

Had there been sufficient interconnector capacity, the TSO would not need to redispatch. Therefore, the cost of redispatching is precisely the cost savings that can be obtained by enlarging the interconnector to the point that there is no congestion. In other words, the cost of redispatching indicates the value of interconnector capacity expansion, when the value of interconnector expansion is taken as the potential savings in operation cost (which is equal to the increase in aggregated consumer and generator benefits). However, redispatching only indicates the momentary value, not the future worth of capacity expansion.

Whether it is economically efficient to actually expand the interconnector depends on the cost of expansion: this should be less than the cost savings. It will therefore not be efficient to expand the interconnector to the point that all congestion is alleviated; at that point surely the marginal cost of expansion is larger than the marginal cost savings from a more efficient redispatch. This means that the economic optimum is characterized by a certain level of congestion. An advantage of redispatching is that the system can be arranged in such a way that the TSO pays both the cost of congestion management and of capacity expansion, so he can balance the latter against the cost of prolonged congestion. He therefore has an incentive to make efficient investments in the network.

Here a choice has been made for a system in which the TSOs do not change the market prices, even though their redispatching actions have in fact changed the marginal generators in both countries. Rather, the financial transactions associated with redispatching are made outside the regular electricity market. The choice for a system in which redispatching is performed completely outside the market was made for two reasons. This method corresponds to the way regular redispatching traditionally is applied within electricity systems. The second reason is that it constitutes a unique case.

A different approach to redispatching is possible, by allowing the redispatching actions to influence the market prices in both countries. This alternative approach starts to resemble market splitting, as different prices develop in the two countries. However, this variation of redispatching will still cost the TSO, as opposed to market splitting which generates revenues (typically for the market operator). The reason for this difference is that the starting point for market splitting is a situation without an interconnector; the market operator consequently gains from trade on the interconnector. With redispatching, on the other hand, the market operator needs to undo some of the gains from trade that were obtained in a simulated situation of abundant interconnector capacity.

A disadvantage of redispatching is that the market does not receive any signals regarding congestion, and will therefore not adjust its trade patterns accordingly. A second disadvantage is that there is a high potential for strategic behavior by the generators. Firstly, the TSO's choice of generators is limited. The solution to the congestion may depend on only a small number or even upon unique generators. Secondly, it may be difficult for the TSO to obtain true cost information from the generators. Traditionally, redispatching took place within vertically integrated companies. In an unbundled system, the generators will try to make a profit by overcharging the TSO or offering less, depending on the case.

A variation of redispatching is counter trading. In this case, the TSO enters the generation market to trade electricity in the opposite direction of the flow which causes the congestion. In theory, counter trading will cost the TSO more than redispatching because the bidding process implies that all generators pay or receive the same amount (De Vries and Hakvoort, 2002b). The advantages are of a practical nature: redispatching requires full knowledge of all marginal cost functions to make economic decisions, while counter trading relies on a bidding process. Therefore it fits better in a liberalized environment and is more transparent. In the case of redispatching, generators will be inclined to inflate or deflate their costs, depending on whether they expect to be called upon to increase or

decrease their output. Thus they enter into a form of strategic behavior, which, in fact is a covert form of bidding. Counter trading makes this bidding process explicit and therefore more transparent. In addition, by allowing most generators a profit, counter trading makes it attractive for generators to be called upon to increase or decrease generation. This may result in a larger pool of available generators to the TSO. Nevertheless, the system of counter trading may also be susceptible to manipulation because a situation may easily develop in which certain generators are indispensable or can only be replaced at high cost.

10.4 Impact of the assumptions

This chapter has presented a theoretical evaluation of several congestion management methods. For the purpose of developing an analytic framework, some strongly simplifying assumptions were made. The most important ones were:

- perfect competition,
- fully inelastic demand,
- stable and predictable prices, and
- a one-dimensional case of congestion; only one transmission link between the two connected markets.

With regard to the assumptions underlying the model of perfect competition, much has already been said. Here the focus is on the latter three assumptions.

Relaxing the assumption of fully inelastic demand does not greatly change the results of the analysis. The congestion management methods of marginal bid explicit auctioning, implicit auctioning and market splitting all have the same effect upon prices, and therefore upon demand. As the prices in country *A* are driven up, demand will decrease somewhat, and vice versa in *B*. Introducing demand elasticity will therefore have the same consequences in each of these systems and does not change the relative merits of each.

The effect of assuming demand to be somewhat elastic rather than fully inelastic will be stronger in the schemes of redispatching and counter trading, as the effect of trade upon the prices in each of the countries is larger. However, the effects upon consumption in the two countries are contrary to each other, so the overall effect will be smaller than the effect in the individual countries. Moreover, demand price-elasticity is quite low in the range of normal electricity prices, so the assumption does not appear to be too strong.

Another assumption was that prices are stable and predictable. In practice, prices can be volatile. This is in part due to the process of liberalization and the associated changes in regulation, but factors like fuel price volatility will always continue to cause uncertainty about future electricity prices. Periods of scarcity of generation capacity also contribute to price volatility (see Chapter 5). The impact of price uncertainty will be larger when a congestion management method is applied in which market parties need to commit to longer periods of time. Price uncertainty can therefore be expected to have a stronger impact upon advance explicit auctions, for instance for month and year-long capacity,

than upon daily auctions. As the uncertainty will tend to depress forward auction prices, it reduces the value of the price signal as an indication of the market value of the link. Implicit auctioning and market splitting function on a daily basis and are therefore less impacted. Unstable prices should not have an effect upon redispatching, as the reimbursements that are made as part of the congestion management method are in principle based upon cost, not upon market prices. With respect to counter trading it depends on the specific details of the implementation of the system. If the TSO has year-long contracts with generators in which the mutual payments are fixed, generators may want some kind of price risk compensation. On the other hand, counter trading may also take place in the (intra) day market.

The final assumption is of a technical nature. The description of the congestion management methods was based upon a two-node radial system: a single interconnector between two markets. Real cases of congestion often involve multiple parallel interconnectors, and often also parallel paths between the two systems through third systems. Thus the one-dimensional models do not reflect the realities of load flow through a network. (See also Section 9.4.) This means that either the congestion management methods need to be adjusted, or they need to be used with large safety margins to account for their inaccuracy. The latter would put a strong claim upon their efficiency, as less interconnector capacity would become available to the market. This is especially an issue for the congestion pricing methods; with redispatching and counter trading, the system operator has easier possibilities to incorporate the realities of the load flow in the management of congestion. The next section discusses a proposal to adjust explicit auctions to the existence of parallel paths with varying degrees of congestion.

10.5 Congestion in a network

A significant shortcoming of conventional explicit auctions is that they ignore the actual load flow pattern of electricity in a meshed network. Explicit auctions are based upon the premise that electricity follows a certain ‘contract path’ between producer and consumer, while in reality perhaps less than half the energy follows this route and the rest is divided among all possible alternative routes. Thus, electricity flows between two neighboring systems may cause flows through other systems, contributing to congestion elsewhere in the interconnected network. This inaccuracy is magnified by the fact that transactions may be linked, as a result of which the contract path becomes a fairly arbitrary construct, which may not even reflect the main flow of electricity (Audouin et al., 2002). The difficulty of controlling the flow of electricity through a network further complicates the matter, as it means that the nominal capacity of parallel lines usually cannot be fully used. The transmission capacity between the two systems is maximized when one of the lines is used to its capacity. Therefore the network capacity that actually is available for auctioning depends upon the load flow, and therefore upon which generators are in use (and also where demand is located, but this is not consider a variable in this analysis).

These inaccuracies of explicit auctions and other contract path related congestion management systems currently are handled by including substantial reserve margins in the calculation of available transmission capacity (ETSO, 2001a). While these reserve

margins are determined from experience, ample capacity should be reserved to minimize the probability of overloading the interconnector (cf. Harvey et al., 1996 and Haubrich et al., 1999). The consequence is that a substantial portion of interconnector capacity is not allocated to the market, so the interconnector is not used efficiently. The simplification of the contract path approach therefore may cause significant inefficiencies in the use of the network. It may even threaten the stability of the system, as was demonstrated in 1999 on July 14th, the French national holiday. Due to lower demand from businesses, France had sold excess electricity to a Swiss trader, who sold it to a German company, who passed it along to another German trader who finally sold it to a Dutch firm. The contract path had a rough U shape, curving around Belgium; a majority of the electricity, however, traveled through Belgium. Without Belgian involvement in the transactions, the Belgian TSO was unprepared for the resulting load flow. Shortly after 9 o'clock in the morning, the Belgian transmission system became overloaded. Only a fast, joint response by the Belgian and French TSOs could prevent the Belgian TSO from needing to disconnect transmission wires to prevent their overheating. This could have caused a cascading black-out through a large part of North-West Europe.

ETSO, the association of European transmission system operators, has since taken measures to improve the dissemination of load flow information among continental European countries, so a near catastrophe like this is not likely to reoccur. However, the fact remains that, due to their contract path approach, explicit auctions use the available network capacity inefficiently. ETSO (2001b) recognizes this and is proposing a system of 'coordinated auctioning' which should improve the technical efficiency of interconnector capacity auctions. The idea behind this system is to refine the technical structure of the auctions, so they better reflect the physical realities of electricity networks, while keeping the commercial side as simple as possible. This way the capacity of the network is used more efficiently, while the market still has the benefits of a simple auction procedure (Chao et al., 2000).

In the ETSO proposal, the location of generators and loads is used to calculate the load flows across the interconnectors. This way the contribution of each transaction to the congestion of the different interconnectors is calculated. Each transaction participates in the capacity auction of each congested interconnector that it uses. An optimization algorithm is used to allocate all available interconnector capacity according to the willingness to pay of the bidders (Audouin et al., 2002). This uses the bid curves of the auction participants and calculates the relative impact of each bid upon the congested links. By accepting bids in such a manner that auction revenues are maximized, it is ascertained that the economic value of the congested links is maximized (similar to a simple explicit auction of line capacity). The users of the interconnector do not experience the complexity of this process; they simply bid for a transaction from one zone to another (not necessarily contiguous ones).

Coordinated auctioning begins to resemble locational marginal pricing in that it uses the exact injection and withdrawal data of loads and consumers to optimize the dispatch of generation within the network constraints. A difference is that network capacity is allocated in a separate transaction, whereas in a system of locational marginal pricing energy and network transactions are bundled. In theory that should not change the

outcome.[67] There are two significant differences between the ETSO proposal and locational marginal pricing. First, the ETSO proposal works with zones, rather than nodes. Calculating a separate price for every node may be unnecessarily complex (Tabors, 1999). Dividing the market into larger zones, separated by congested links, may be just as effective, while it would be much more simple and transparent.

The second difference between coordinated auctioning and locational marginal pricing is that the latter uses the data from all generators, while coordinated auctioning uses only the data from the generators that participate in inter-zonal transactions. The data from generators who do not export are only included indirectly, as each involved TSO needs to compile a load flow forecast for his system. This has two significant consequences. One is that, because not all data is available to calculate the entire load flow, the calculations of the flows over the interconnectors remain approximations, so there still is a need for a safety margin. Coordinated auctions improve the technical efficiency but do not maximize it. The second consequence is potentially more damaging. A generation firm may have some generators who produce for consumers within their own zone and other generators who export to other zones. Such a firm may be able to impact the occurrence of congestion by shifting the output that is nominated for export between its generators, as generators in one location may contribute more to congestion than other generators (Boucher and Smeers, 2002). This may be used to artificially increase or decrease the congestion, depending on the firm's competitive relations. Thus, the firm may be able to impact the export opportunities of its rivals and possibly even the price difference over the interconnector.

Nevertheless, coordinated auctioning could be a step forward for European markets, as it increases the efficiency with which the interconnectors are used. It leaves the market within the zones unaffected, so all opportunities for bilateral contracts and trade within electricity exchanges remain the same, while the requirements for inter-zonal trade are kept to a minimum. An important aspect for Europe is that coordinated auctioning, like 'simple' explicit auctioning, functions fairly independently of regular transmission network and system operation. This means that it can be used to manage the congestion on interconnectors between differently structured electricity systems, just like 'contract path' explicit auctions.

Coordinated auctions refine the locational incentives provided by conventional explicit auctions. The electricity market price difference that the auctions create provides an efficient locational signal to generators, provided that the price difference is the result of real differences in the marginal costs of generation (and not artificially created, for instance through differences in taxes, fuel subsidies, environmental regulations or through manipulation of the auctions by bidders). By signaling regional differences in the marginal cost of electricity generation, coordinated auctions provide generators with an incentive to invest in locations where their value is highest, thereby reducing network

[67] Chao et al. (2000) also propose a system of co-ordinated auctions ('flow-based transmission rights') for the California market that is based upon a mandatory pool in which the pool operator also clears the spot market and takes care of congestion management.

congestion. However, they are not perfect: they are complex, they do not use the network optimally and they can be manipulated.

10.6 Comparison of the congestion management methods

10.6.1 Welfare effects

How much consumers pay under the different congestion management regimes depends, first of all, on the actual costs of generation and transmission and, secondly, on the distributive effects, in particular on the profits that are made by the generators and the TSOs. In the short term the system costs are the same for the reviewed methods. This means that the net social benefit is the same for all these methods in the short term. As the revenues of generators and the TSOs equal the payments by consumers, one party's gain means another party's loss. The general effects are similar for the congestion pricing methods on the one hand and for the corrective methods on the other.

One cannot draw general conclusions regarding the welfare effects from this analysis. In particular, one cannot assume that congestion pricing methods will cost consumers more than the corrective methods. The occurrence of congestion rents may cause the congestion pricing methods to appear more costly but this is not necessarily true. Whether consumers benefit more from the corrective methods than from the congestion pricing methods depends entirely on the manner in which the congestion rents from auctions or market splitting are spent, respectively on how the money to fund congestion management is raised in the case of redispatching and counter trading. The apparent higher cost of the congestion pricing methods may be compensated, from a consumers' point of view, by spending the congestion rent on something that consumers would otherwise have to pay. The rents can be used to reduce transmission prices or for capacity expansion projects that otherwise would be financed by the network users. The operational cost of the corrective methods, on the other hand, will need to be passed on to the consumers in one way or another, for instance through inclusion in the transmission tariffs.

It follows that undesired welfare effects of congestion management methods can be compensated elsewhere, for instance through adjustment of the general transmission tariffs. The methods may differ with respect to how easily this can be done, however. In the form that was chosen here, redispatching and counter trading may leave the generators with higher profits than the other options, and it may be difficult to return these profits to consumers. The rents from the congestion pricing methods can be used more easily to the benefit of consumers because they are available as a separate revenue stream.[68]

[68] The EC proposes to limit the possible applications of congestion rents, see the proposed Regulation Of The European Parliament And Of The Council on conditions for access to the network for cross-border exchanges in electricity, Art. 6.6.

10.6.2 Economic efficiency

All the congestion management methods that were analyzed in this chapter are in theory capable of achieving the least-cost dispatch of generation. The generators that are allowed to export via the interconnector are in each case selected upon their marginal production cost. As the interconnector capacity is given to the cheapest available generators, this means the overall cost of generation is minimized. Therefore all reviewed methods are economically efficient in the short term.

Explicit auctions are likely to create higher transaction costs than the other methods, as they require two separate transactions for cross-border trade of electricity, whereas the other methods require market parties only to make a single transaction. The complexity of auctions may pose a barrier to market parties, in particular to those who do not have the expertise to handle the associated risk. As a result, explicit auctioning may have a distorting effect upon the market. Small generators may be least favored, as for them the transaction cost of the auction is relatively highest. On the other hand, a well-designed auction may improve the transparency of the market, which would reduce the barrier for newcomers and stimulate competition.

The other methods appear to have lower transaction costs than auctions. However, this may be offset by their lower transparency. Implicit auctions require market parties to bid into a different organized market than in their own region. The associated transaction costs should be small. In the case of market splitting, the transaction costs of trades across a congested link are not different from those of trades that do not involve a congested link, so the congestion management system does not impose an extra barrier to trade at all. Redispatching and counter-trading also pose no barriers to trade, as the market operates as if there is no congestion but they do involve extra costs for the TSO. A disadvantage of redispatching and counter trading is that, depending on the situation, they may be quite susceptible to strategic manipulation by generators.

A complication arises when sales of electricity are made across two consecutive congested links. This may pose a significant obstacle to trade, unless the congestion management methods are easy to handle for market players. The conclusions with respect to transaction costs apply a fortiori. If capacity auctions are used, combinatory bidding must be implemented to allow transactions involving the multiple congested links. The ETSO proposal for coordinated auctioning provides this feature but the simple explicit auctions as they exist now do not. Market spitting should also be able to accommodate trade between non-contiguous zones. Redispatching and counter trading would, of course, still pose no obstacle to trade. Implicit auctioning and market splitting could work but would require a significant level of organization between all the involved TSOs and market operators.

The main shortcoming of explicit and implicit auctions is that they are ‘contract-path’ solutions: the ‘route’ that the contracts follow is not the same as the route that the electricity takes through a meshed grid. As a result, auctioned capacity may not be fully used or, worse, may not actually be available. To avoid the latter situation, it is necessary to use substantial safety margins in the calculation of the available transmission capacity. This reduces the amount of capacity that is available to the network and therefore the

economic efficiency. Again, the ETSO proposal coordinated auctions, promises to remediate this shortcoming. Because market splitting requires all production and consumption to take place within the local zone, it is not a contract path solution. The market operator can optimize the flows between the zones within the network constraints.

A final issue is the netting of counter flows. The issue arises when there are transactions in opposite directions across a congested link. Transactions in one direction contribute to the congestion; due to the fact that electricity flows in opposite directions cancel each other out, transactions in the opposite direction reduce the congestion. Congestion management method should reward counter flows for their positive contribution. A second reason is that counter flows, if valued properly, can reduce the economic inefficiencies caused by oligopolistic pricing by generating companies (Hobbs et al., 2004). The different congestion management methods differ with respect to whether they can accommodate counter flows in this manner. Redispatching and counter trading correct congestion that exists after all transactions have been processed, so they automatically incorporate counter flows.

For the congestion pricing methods, the key is whether the market parties only pay for a *right* to use capacity, or whether they also have an *obligation* to use that capacity. When they have a right, but no obligation, the TSO does not know for sure whether the scheduled flows will actually take place. Then it is risky to assume that counter flows will reduce the congestion. Counter flows are often not netted in explicit auctions, even if they apply the 'use-it-or-lose it' principle. The separation from transmission rights and power flows apparently makes it too difficult to estimate in advance the extent to which the capacity actually will be used. Implicit auctioning is, by its nature, a one-directional way of congestion management. Market splitting, on the other hand, does provide the possibility of netting counter flows, as it is based upon firm bids.

Explicit auctions are a popular option for managing congested interconnectors in Europe, at least in the near term. The main exception is the Nordic market, where market splitting is used. While deemed an elegant and efficient procedure, the current model of market splitting would need to be adjusted in a manner similar to the coordinated auctions to be able to function in the meshed network of mainland Europe. However, as then it would start to resemble locational marginal, similar obstacles to its implementation arise: the required levels of harmonization and market integration would be high. Explicit auctions are more robust in this respect: they can be applied to systems with entirely different regulatory structures, network access rules and transmission tariff systems. This, combined with their transparency, currently makes them the favorite option for managing congestion within the EU.

10.6.3 Long-term signals

Ideally, a congestion management method is not only economically efficient in the short term but also provides efficient long-term incentives to both the network managers and to generation companies (EC, 2000). None of the reviewed methods combines these goals. Congestion pricing systems signal the cost of congestion to market parties, who may adjust their investment behavior accordingly if the congestion persists. Thus they will, in

the long run, reduce congestion. However, congestion pricing methods provide no incentive to network managers who have a monopoly for relieving congestion. If the network manager is allowed to keep the congestion rents, the incentive is even to increase congestion. Nevertheless, congestion pricing methods yield a revenue flow, which can be dedicated towards projects to relieve the congestion.

The corrective methods have precisely the opposite effect: while they do not provide market parties with an incentive to change their behavior to reduce congestion, they do provide network managers with an efficient incentive to minimize congestion. If network investment were competitive, congestion pricing provides, at least in theory, optimal incentives for network expansion (cf. Nasser, 1998; Joskow and Tirole, 2003).

A choice needs to be made between providing efficient incentives to the generation market or to network managers. The price difference that is created by congestion pricing induces generators to locate in the high-priced zone, even though the cost of generating is higher there. By doing so, the demand for imports is lowered and transmission costs are reduced, so general system costs are lowered. When redispatching and counter trading are used, on the other hand, market parties are confronted with the costs of their transactions. They pay standard network tariffs which means that the particular costs of their transactions, if they exceed the standard tariffs, are socialized. Therefore they may sell electricity across the congested interconnector even if the price advantage is smaller than the congestion cost. It is true that at the operational level redispatching compensates this and should lead to an optimal dispatch of generation. However, because the congestion costs are socialized, the market receives no incentives for relocating generation facilities in a way that reduces system cost. In the long term the system costs will therefore be higher than when congestion pricing is used.

Price signals from congestion management methods provide one of only a few options for influencing generator investment behavior, as was seen in Chapter 9. As monopolists, TSOs, on the other hand, are inherently subject to more extensive regulation, which provides possibilities to include a process for efficient expansion of interconnector capacity. Therefore it seems to be more important to provide efficient investment incentives to market parties than to TSOs. By creating a price difference across the congestion, the congestion pricing methods signal the value of new generation capacity in the constrained area.

While the signal from auctions is not very specific – the 'constrained area' may be as big as a country – at least it provides a rough indication of the geographical differences in the value of generation capacity. The price difference created by these congestion management methods therefore contributes to the long-term economic efficiency of the system. Thus, the fact that it reduces some opportunities for trade should not be considered a disadvantage (as the EC apparently regards it) but as a plus (EC, 2001b). The price difference prevents some of the most inefficient market outcomes, for instance due to consumers purchasing electricity from remote power plants for a price advantage that is smaller than the additional transmission cost. Explicit and implicit auctions unfortunately are too onerous to apply with a high resolution, so their application is limited to the major bottlenecks. They could be used on some congested links within

certain large countries but they would not be an efficient solution for more small-scale applications. Market splitting is more versatile, as for the market parties the bidding process is not impacted by the occurrence of congestion. The management of congestion is performed by the market operator after all the bids have been received. The market operator can therefore merge or split zones at will, so he can respond quickly to the development of new bottlenecks.

10.7 Conclusions

This chapter analyzed four market-based congestion management methods and modeled their theoretical short-term economic impacts: explicit auctioning, implicit auctioning, market splitting and redispatching (and counter trading, which is a variation of redispatching). In theory, all four congestion management methods can achieve short-term economic efficiency in the sense that they lead to the most efficient dispatch of the generators, given transmission constraints. Their differences lie in the distribution of costs, their practical feasibility and their long-term impacts.

The reviewed congestion management methods can be divided into two categories. Explicit auctioning, implicit auctioning and market splitting are forms of congestion pricing: they regulate access to the congested interconnector through some form of a price mechanism. Redispatching is a corrective method: it allows market parties to act as if there were no congestion but require the TSOs to take corrective measures. Redispatching (and its variant counter trading) costs the TSO money which he may be able to recoup through transmission tariffs. The overall distributive effects of the reviewed methods cannot be assessed as they are dependent on how the costs of redispatching are recovered and how congestion rents from the congestion pricing methods are used.

The corrective methods have as an advantage that they are the simplest options to work with for market parties. The corrective methods do create extra transaction costs for the TSOs, however. An advantage of the corrective methods is their versatility and the short time frame within which they can be applied. Therefore redispatching and counter trading are useful as back-up options for unexpected cases of congestion and for fine-tuning other congestion management methods. They also provide an efficient signal to the TSOs regarding the demand for network capacity expansion. The main drawback of the corrective measures is that they do not provide market parties with an indication of the cost of congestion. This is a significant point because they do not provide generators with an incentive to adjust their long-term (investment) behavior.

The congestion pricing methods create the opposite long term incentives. They provide an efficient signal to generators regarding the cost of using the congested link but they do not provide the TSOs with an incentive for optimal capacity expansion. The objective of providing both market parties and TSOs with economically efficient incentives appears impossible to reach. Facing a choice, congestion management methods that provide market parties with incentives for efficient long-term behavior must be preferred. To compensate for the lack of incentives, it is probably easier to influence the network

planning process, as it is part of a regulated monopoly, than the investment decisions of generation companies.

The main drawback of the congestion pricing methods consists of their higher institutional requirements. In addition, the conventional system of explicit auctioning uses a contract path approach, which is inefficient. An improvement, called coordinated auctioning, is being developed by ETSO. By including the actual load flows in the calculation of available capacity, the efficiency with which the network is used should be improved.

10.8 Recommendations

10.8.1 Policy recommendations

In cases of structural congestion, congestion pricing methods are preferable to corrective methods because they give better incentives to the generation market. Explicit auctions are a good starting point but their transaction costs give reason to consider market splitting in the long term. However, this method must first be developed for a meshed network.

A number of reasons have been identified why private investment in interconnectors would lead to a sub-optimal volume of interconnector capacity. As the design and operation of an interconnector also impacts the operation of the connected networks, it should probably be considered part of the network monopoly. An exception may be made for DC links, which are not subject to the same externalities as AC links. The higher cost of DC interconnectors may be a reason to seek private sector investment.

Interconnectors often link electricity systems with different market rules, tariffs, taxes and subsidies. As a result, the flows across interconnectors are not necessarily the result of intrinsic economic factors but a product of the differences in regulations. The resulting demand for transmission capacity is therefore not necessarily economically efficient and is subject to change when the regulations in question change. Care must be given not to base decisions regarding investment in interconnectors upon these artificial flows.

10.8.2 Research recommendations

Currently, a system of coordinated explicit auctions is being developed that should mitigate a number of the short-comings of one-dimensional auctions. Coordinated auctions use available network capacity substantially more efficiently than traditional one-dimensional auctions. However, they do not remove the disadvantage of needing two transactions to transmit electricity across an interconnector. Market splitting does solve this issue and therefore has substantially lower transaction costs and risk to traders. Therefore the possibility of developing a multi-lateral system of market splitting should be explored, as it could form an elegant congestion management method for European interconnectors.

11 Synthesis and reflection

The lessons of the previous chapters are drawn together in this chapter and the common causes and policy implications of the generation adequacy and coordination issues are discussed. The approach that was used and the impacts of the assumptions that were made in this study are reviewed. At the end of the chapter, the scope widens and the limits of the competitive paradigm and the implication of the findings for other sectors are considered.

11.1 Introduction

In the preceding chapters, two types of market failure in electricity generation were investigated: the risk that the total volume of generation capacity will be insufficient and the risk that the development of the generating stock is not sufficiently coordinated with the network. This chapter starts with a synthesis of these issues, considering their parallels and differences, and then zooms out to make more general reflections. The next section reviews the common physical features between the issues of generation adequacy and coordination of generation investment with network development. Next, in Section 11.3, common policy implications are discussed. The last two sections of this chapter are dedicated to more general lessons. Section 11.4 provides a reflection upon the methodology of this study and the assumptions that were made. Section 11.5 reflects upon the limitations to introducing competition in technically and organizationally complex sectors. Section 11.6, finally, draws some lessons for other sectors.

11.2 Common physical features

11.2.1 Network externalities

Network externalities hamper the creation of efficient investment incentives both for the purpose of generation adequacy and for coordination of generation with the network. First, the network itself has a public good character because the costs of specific transactions cannot be allocated unambiguously to the network users. Second, all electricity consumers who are connected to the same network experience the same reliability, at least as far as it is determined by generation adequacy, so electricity

generation also has a certain public good character. This aspect can be changed, however; for instance through a system of capacity subscriptions.

Attention for both the adequacy and the coordination issues appears to be slowly emerging in Europe. In the USA, both issues are more widely recognized. Generation adequacy is most commonly addressed by using capacity requirements, while locational marginal pricing is intended to provide the correct locational incentives. Two myths may be blamed for the fact that these issues have received little attention in Europe. The myth of the invisible hand suggests that supply and demand will always be balanced through the price mechanism, which Chapter 5 showed not necessarily to be true in electricity markets. The coordination issue is masked by the myth of the copper plate, which suggests that the European electricity networks (should) function as a copper plate, where anyone can inject or withdraw any amount of electricity at any location without limitation.

Both myths appear related to an overly strong emphasis on, and belief in, competition as the means of improving the efficiency of the electricity sector. While competition undoubtedly puts pressure on the generation market to increase efficiency, it may not necessarily lead to an optimal dynamic development of the market. Similarly, while the electricity network needs to facilitate competition in the generation market, the need for coordination may not be neglected, at the risk of creating in efficiencies (rather than reducing them) or reducing reliability.

11.2.2 Differences in time constants

A second commonality among generation adequacy and coordination is the effect of the differences in time constants between different parts of the system. The market has the shortest time constant, with significant variations occurring on a daily basis. Generation facilities have an expected economic life on the order of two decades. Therefore, generation stock is necessarily slow to react to the market. Network assets, finally, often have an even longer life span, ranging up to more than five decades. Moreover, network development has a certain path dependency, as changes to the basic structure of the network are orders of magnitude more costly than replacement of old or failed components. Therefore, the time constant of the network is much longer than the economic life of its components. The inevitable slowness in the way that the network adjusts to changes in the generating stock is substantially increased in many countries by the length of the permitting process for above-ground power lines. The differences in time constants pose a coordination problem with respect to the long-term development of a decentralized electricity system.

The electricity market versus investment in generation facilities

The difference in the time frame within which the market operates and the life cycle of generation facilities increases significantly the risk of investment cycles. The relative slowness with which generation stock can be adjusted to changes in demand provides a double obstacle. First, it gives reason for generating companies to be cautious when investing in capital-intensive generation facilities. Second, it means that when there is a

shortage of generation capacity, supply cannot be augmented quickly.[69] The effect of different time constants is exacerbated by the volatility of electricity prices and the limited degree to which forward markets develop. Uncertainty about future prices and the absence of sufficient hedging tools are reasons for generating companies to discount expected future prices, which leads to a lower level of investment in generation capacity.

A break-through of distributed generation technology could fundamentally alter the dynamics of the electricity market. If small-scale generation facilities could be produced in series and be ordered within a few months or even less, the difference in time constants between the electricity market and the generation facilities market would decrease substantially. As a result, the risk of investment cycles would also diminish. If the electricity generation units are small enough to be mobile, investment risk would decrease further, as excess capacity in one location could be moved elsewhere. Distributed generation technologies that fit this description, such as micro-turbines and fuel cells, currently are being developed but have not entered commercial use widely (Dondi et al. 2002).

For the moment, however, the difference in time constants means that it is unlikely that the market will provide a socially optimal level of generation capacity. In the presence of uncertainty, generating companies will want to avoid excess capacity, as they cannot recover its cost (at least in a competitive market). If the market 'undershoots', on the other hand, there will be a shortage of generation capacity, accompanied by high social costs. Chapter 5 argued that from the perspective of society, it is wise to overinvest to some extent. The capacity mechanisms that were described in Chapters 6 through 8 can be used for this purpose.

Generation versus the network

Investors in generation facilities face uncertain signals from market parties while network development is complicated by its slowness relative to the generation market. At an operational level, changes in the load flow are restricted by the locations of generators. However, these changes can still be significant, as is the experience in western Europe. There, many interconnectors between national systems have become congested since liberalization, so the TSOs are faced with the question of whether and where to expand capacity.

With respect to investment, network managers may face difficult decisions. The first obstacle to the coordination of the development of network capacity with the generation market is the difficulty of obtaining permits for above-ground power lines. In addition,

[69] During much of the history of electricity systems this was, for several reasons, not an issue. One is that in the vertically integrated monopolies before liberalization, the financial risk to the utility of over-investment often was limited, depending on how the utility was regulated. In many European countries, this resulted in a bias towards large capacity margins. Secondly, long-term planning of generation capacity is easier for a regional monopoly that only needs to consider the development of total electricity demand, than for a competitive generating company that also needs to consider its market share. Finally, electricity consumption grew so fast during much of the twentieth century that any over-investment in generation capacity soon was absorbed by the growth in demand.

under incentive regulation network managers may have reason to be risk averse and invest only when it is clear that there is a lasting demand for more capacity. Because generation facilities may be constructed and retired in less time than necessary network adjustments can be made, there is a risk that network development permanently lags behind the needs of the electricity market. Even perfect incentives to network managers may not necessarily lead to socially optimal network development. Typically, incentives are not optimal; for instance, in many cases the costs of scarcity of network capacity are external to the network managers.

As with investment in generation capacity, an asymmetry of risk may develop. For society, it is desirable to have excess network capacity for the sake of security of supply and to stimulate competition. A mechanism is needed which stimulates network managers to develop robust networks, without forfeiting incentives for efficient operating and investment decisions. Under rate-of-return regulation the risk of underinvestment would be smaller but there would be a risk of inefficient investment.

A modular solution appears less feasible than in the case of generation but a break-through of distributed generation could simplify network development. While it would require significant adjustments to the networks in the transition phase – especially the way they are operated – distributed generation would both lower average network costs and investment risk (Kirby and Hirst, 2000). The costs would be lower because less network capacity (per unit of electricity consumed) would be needed, as electricity would be produced closer to consumers. Network investment risk would be lower because generation would take place close to loads, which in the aggregate are not as likely to change in location as large generation facilities (cf. Poza and Ackermann, 2001; Dondi et al., 2002).

Different time constants and the effectiveness of financial incentives

Liberalization rests upon the principle of decentralization of decision power. Actors in the field inevitably have more detailed, accurate and up-to-date information than planners at the system level do. If they operate under incentives to act in the general interest, they should be able to make better decisions than system planners. An important goal in designing a market is to provide all agents with efficient economic incentives. This study has shown that there are many obstacles to developing efficient incentives in the electricity sector due to network externalities and due to the different time constants of the electricity market, the generation investment cycle and network development. Even if the externalities could be internalized in a perfect system of incentives, the differences in time constants presents an obstacle to a socially optimal outcome.

The use of financial incentives is grounded in neo-classical economic theory, which is based upon an equilibrium model of markets. It is presumed that the price mechanism always leads to an equilibrium between supply and demand: if there is a shortage of a product, higher prices lead to a higher production rate and a lower consumption rate until the two are in balance and vice versa. Time is not considered as a variable: it is presumed that the presence of a feed-back loop is sufficient for an equilibrium to develop. This is not necessarily the case, however: if the supply side cannot react fast enough to the price

signal, the system may begin to oscillate, leading to a pattern of investment cycles.

In the electricity sector, investment cycles may develop as a result of the short period of time between the first development of price spikes and a situation in which there is insufficient capacity to meet demand. As was argued in Chapter 5, planning and developing new generation capacity takes so long that it is likely to come too late to avoid a prolonged period of scarcity, if investment decisions are made in response to price spikes. Chapter 9 made a similar argument with respect to the networks, except that a permanent development lag is the more likely result if a system of incentive regulation is applied to the networks. Thus, the analysis of the generation adequacy issue and of the coordination issue leads to the conclusion that the existence of efficient economic incentives is not sufficient for an efficient long-term development of the system.

11.3 Common policy issues

Robustness versus economic efficiency

The conclusion of the previous section places the pursuit of efficient economic incentives in perspective. If even an optimal incentive scheme does not lead to a socially optimal outcome, perhaps it is not such an important a goal. Chapter 5 argued the case for additional measures to secure generation adequacy, which were subsequently discussed in Chapters 6 through 8. These are intended to ensure a socially acceptable outcome even in the presence of market imperfections. A similar approach to network development will also serve society's interests better than trying to manage the complex relations between the networks and the electricity market only through economic incentives. A limited degree of overinvestment may therefore be pursued as a form of social insurance against the much higher social costs of service disruptions.

The trade-off between the pursuit of economic efficiency and robustness becomes clear from the analysis of the issue of generation adequacy. An energy-only market is optimally efficient in theory but is susceptible to investment cycles if demand is not sufficiently involved. The most attractive alternatives within the current physical system, capacity requirements and reliability contracts, would reduce the risk of capacity shortages although likely at the price of a certain amount of excess capacity. The trade-off can be avoided through the introduction of capacity subscriptions, as they let individual consumers choose the volume of generation capacity themselves.

The same trade-off is more difficult to avoid with respect to the networks. Future load patterns are much less certain since liberalization, as the location and output of generators can no longer be planned. Aiming for economically efficient development contains a risk of substantial deviations from the optimum, including periods of inadequate network capacity. This could be avoided by overinvesting, with the goal of economic efficiency forfeited, to a degree. Again technological innovation may provide a way out: a shift towards distributed generation might reduce or even solve this issue. This is an uncertain scenario, however: the necessary technology has not yet made a commercial break-through and it is unclear what the exact system dynamics would be in a distributed

electricity supply industry.

Responsibilities

In the European model responsibilities are not clearly defined for either the issue of adequacy or that of coordination. This raises the question of where the responsibility for these functions should lie. With respect to generation adequacy, responsibilities need to be assigned for:

- deciding the level of reliability,
- operating the capacity mechanism, and
- monitoring its effects.

For the coordination issue, the fact that a combination of solutions will need to be used means that responsibilities for the issue probably will be distributed among different agents.

In principle, it is the task of the network managers to adjust their networks optimally to the demand for transmission services. However, in the European system with fixed network tariffs, they have no strong incentive to do so except when congestion is managed through redispatching or counter trading. These methods provide network managers with an economically optimal incentive for network expansion. However, the trade-off is that they do not provide efficient incentives to the generation market, so they only affect one aspect of coordination. Chapter 10 argued that, given the need to choose, it is more important to provide efficient economic incentives to the generation market than to network managers, as network development can be regulated more easily. Thus, congestion pricing methods are favored.

This leaves the question of how to stimulate network managers to develop their networks in an optimal manner unresolved. There are two basic options: the creation of proxy incentives and 'command-and-control' regulation. Proxy incentives, such as bonuses, can stimulate short-term efficiency, for instance by reducing congestion. Direct regulation does not stimulate efficient behavior by the network manager with respect to minimizing congestion but could be necessary for network expansion. In this case, the responsibility for network development would be shared between the network company and the regulator.

With respect to the need to adjust generation decisions – investment, retirement and operation – to the network, congestion management may be used to provide incentives to the generation market, as was seen in Chapter 10. Congestion pricing methods provide geographically rough but generally efficient incentives. It could be supplemented with variations in network access charges. The network manager could also be given direct authority to influence generator siting decisions through financial incentives (Stoft, 1999). If all else fails, siting decisions could be left to a process of regulatory approval that includes a system-wide benefit-cost analysis. This would, again, shift some of the responsibility for system development to the regulator.

Scale and timing of the issues

The issues of generation adequacy and coordination are at different stages in the policy process. While the evidence is mounting that the generation market needs to be adjusted to provide adequate investment incentives to generators, the full scale and scope of the coordination issue remain to be investigated. In the area of congestion management, policy intervention has already been necessary but other aspects of the coordination issue remain unexplored. The implications for public policy vary. The need to take measures to protect generation adequacy in Europe is becoming urgent, while the issue of coordination requires more research with respect to the need to adjust the market design.

Data collection

For both generation adequacy and coordination more data should be collected but for different reasons. For coordination, better data are required to establish the scale of the issue. For instance the degree to which international trade in Europe is distorted by differences in taxes and subsidies, and the degree to which network charges give rise to inefficient location decisions of generators. With respect to adequacy, monitoring the market is not a useful means of verifying the problem because observation of current market trends will not provide a timely warning. The time delay between the decision to implement a capacity mechanism and a resulting increase in investment in generation capacity is too long; once an unacceptable decrease of the generation margin is observed, it is too late to intervene. Monitoring the market may, however, serve to verify the effectiveness and efficiency of a capacity mechanism. Van Werven (2003) proposed a framework for monitoring generation adequacy.

11.4 Reflection upon the method and assumptions

11.4.1 Method

Analysis

The qualitative analysis of the issue of generation adequacy provided arguments for market intervention. While the future cannot be predicted, and the long-term development of markets can therefore not be anticipated with certainty, the arguments as to why a market would not produce sufficient capacity provide reasons for market intervention. A more quantitative analysis of the question of generation adequacy may be performed but would not be able to provide a definitive answer. Too many fundamental factors cannot be quantified (such as regulatory uncertainty, the investment strategy of an oligopoly, the degree of risk aversion of investors, the effects of incomplete information) for a quantitative model to be sufficiently complete to provide an answer. A model is as good as its assumptions, and in this case many assumptions would need to be made that would be difficult to verify but that would have a significant impact upon the outcome of the model. An example is provided by Visudhiphan et al. (2001), whose model shows that investment decisions based upon historic spot market data leads to a higher deficit than if they are based upon forward contract prices. Thus, the assumption about the type of information that investors use is crucial, at least in the perfectly competitive setting of

the model.

The focus of the argument in this study was not limited to the question of how to achieve a narrow social investment optimum but included consideration for the situation in which an optimum would never be reached in a dynamically developing system. The asymmetry of the loss of welfare function is highly relevant in this perspective. This is another function that cannot be quantified easily, especially for a number of years into the future. However, minimizing the loss of welfare from market failure is precisely the issue. Good public policy should not only be designed to achieve an optimal outcome but also to be robust against unforeseen events. A quantitative analysis necessarily emphasizes the former and risks disregarding the 'what if' scenarios. These can be included in a quantitative analysis as well but, due to the many assumptions, the end results are not necessarily more convincing than those of the analysis in this study. The qualitative approach used in this study allows consideration for these different perspectives in a much less onerous manner.

The advantage of a qualitative analysis is that the argument is more accessible to people without a technical background. Scientifically, this means that verification of the results is open to more people than in the case of a quantitative model. Accessibility of the argumentation is important for the social relevance of the study, as the final aim is to provide policy advice.

An important aspect of the evaluation of capacity mechanisms is their dynamic behavior. Therefore, this part of the analysis is supported by a dynamic model (in the Appendix) that provides an indication of the stability of several capacity mechanisms. The model inevitably uses some simplifying assumptions. An important one was that it does not consider the effect of market power upon electricity prices or strategic investment behavior. Therefore, the model must not be used to forecast future developments; the development of an investment cycle in many runs of the model cannot be interpreted as any kind of proof that this will happen in reality. However, the model does provide a framework for comparing the degree to which different capacity mechanisms are able to stabilize investment in a market that otherwise would be subject to investment cycles.

The situation is different with respect to the coordination issue. The qualitative analysis provided a first step, a structuring of the issue, but does not provide an indication of its severity. In the case of congestion management, the need for additional instruments is evident. With respect to other issues like reactive power management or the locational decisions of generators the need for intervention is not clear. Quantitative modeling and empirical research (for as far as empirical evidence has emerged) can provide an indication of the extent of these issues. As solutions, other than locational marginal pricing, would not be theoretically efficient but have a pragmatic nature instead, they also would need to be based upon a quantitative calculation of their impacts.

Whereas a qualitative problem analysis was used for both the adequacy and the coordination issues, in both cases policy options were modeled quantitatively. The evaluation of capacity mechanisms is supported by a system dynamics model (in the Appendix). The model results support the results of the qualitative evaluation in Chapter

7. The analysis of congestion management methods for European interconnectors was supported by a simple economic model which had as a goal to determine the incentives they provided and their potential efficiency. This approach led to a useful first categorization of solutions, which showed which ones were more suitable in a liberalized environment. The final choice of method will depend upon more practical and less quantifiable criteria, such as feasibility in a meshed network and transaction costs.

Conceptual framework

The most basic assumption underlying this analysis is that the economic and institutional organization of the sector must reflect the technical requirements, e.g. the need to balance the system or to coordinate network and generator operation. A conceptual framework was developed to analyze these relationships. While the formal framework was not used everywhere, the basic approach of starting with the technical requirements of the system and deducing from them the requirements for the economic organization of the sector was used throughout. This is a fruitful approach, which exposes gaps and inconsistencies in the current market design.

Taking the relations between the technical and economic subsystems as a starting point avoids both economic dogmas (such as the notion that by definition a market provides an optimal level of investment) and technical dogmas. An example of the latter is the need for a reserve of generation capacity for the purpose of long-term reliability. The approach here shows that the very notion of such a reserve is tied to the pre-liberalization paradigm. In a market, there is a range of generators with different characteristics which determine their merit order. The question therefore is not how to create a reserve but how to design the system so all generating companies together produce an optimal volume of generation capacity. Rephrasing the question thus opens the door to more innovative and market-oriented solutions.

11.4.2 Impact of the assumptions

Focus on long-term issues

This study focused almost entirely upon long-term issues. The exception is the abuse of market power in the form of capacity withholding by generating companies during periods of scarcity. This exception was made because the incentives for withholding generation capacity are a result of the market structure. The California crisis demonstrated that the abuse of market power can contribute to an electricity shortage in a highly damaging way. Therefore, any adjustment to the market structure for the sake of securing generation adequacy must also take into account the possible abuse of market power. Implicitly, the assumption was made that the best way to avoid capacity withholding is to change the incentives to generating companies so they do not benefit from it, rather than trying to suppress it through, for instance, legal action. Again, the California crisis provides evidence on the difficulty of mitigating market power through competition law. Prevention – through a change of the operational incentives to generating companies – would be less costly and more effective. Changing incentives is a matter of market design, the subject of this study.

There are also other short-term issues with an impact upon the reliability of service, such as system security (resilience against sudden, large disturbances), maintenance planning and operational stability. As these issues are independent from the market design issues that are studied here (except, of course, that the more capacity that is available, the easier it is to deal with these issues), they were not included. Regarding operational stability, an effect of liberalization is that the responsibility is spread among more agents, which gives rise to the concept of networked reliability (Roe et al., 2002).

Hydropower

The analysis was made for systems without a dominant role by hydropower. The presence of storable hydropower changes the analysis of both the adequacy and the coordination issues substantially. Hydropower-based systems typically have far more installed generation capacity than they need in order to meet peak demand. Their output is constrained by the volume of water in the reservoirs: they are energy-constrained, rather than capacity constrained. The uncertainty in hydro-based systems is even larger than in other systems, as the supply is subject to significant annual variations in precipitation.

When a hydropower-based system is short of generation capacity, the system would need a new generating plant to augment the average output. In most countries, all available opportunities for hydropower have already been developed so the option that typically is considered is a base-load nuclear or fossil fuel plant. In a competitive setting, the difficulty is that during wet years this plant could not compete. Moreover, it should operate even when the reservoirs are relatively full to minimize the risk that they would be drawn down too far. In some cases, wind power may provide a solution as the marginal costs are almost zero, like those of hydropower. As its operating costs are minimal, it would always be in merit. Its short-term fluctuations could be compensated by the hydropower plants, while it would allow a lower average production of hydropower.

The issue of coordination, on the other hand, is greatly simplified by the presence of hydropower. Hydropower plants cannot be moved and are not likely to be closed unexpectedly.

No attention for generator location in the analysis of generation adequacy

The issue of generation adequacy was simplified by disregarding the impact of the location of generators. Network constraints may reduce reliability even if the overall volume of available generation capacity is sufficient. This was the case in San Francisco during the first outage of the California crisis in the summer of 2000. This issue does not change the analysis made in this study but adds another dimension to it. The need for an adequate overall volume of generation capacity remains; however, a second requirement is that the network can accommodate the likely load-flow in different contingency scenarios. This latter aspect can only be determined through load-flow modeling of specific networks. It is not a consequence of liberalization, however, so more experience exists regarding its analysis.

Transmission tariffs that are not value-reflective

The analysis of the coordination issue was made for decentralized electricity systems in which transmission tariffs are fixed. As far as the author knows, the only value-reflective system of transmission pricing is locational marginal pricing, which appears only possible in an integrated system. Transmission tariffs that are not value-reflective create external costs and benefits (except for a system of *ex post* network tariffs, which is presumed to be unacceptable to market players). Therefore, all systems that are not based upon locational marginal pricing are susceptible to the kinds of issues that were discussed in Chapter 9.

Fuel security

The issue of fuel security was not addressed in this study. One reason is that it has been recognized as an issue at least since the first oil crisis, so it has been researched well. Liberalization of the electricity market has not changed the issue. Liberalization of the fuel markets themselves may affect the security of supply of these fuels and therefore also of electricity, but this is a question of the liberalization of these fuel markets, not of the electricity market.

Operational constraints

Operational constraints upon generators were not considered. A well-known issue is that during hot weather, cooling water regulations may limit generator operation. Other environmental restrictions may apply as well. These limitations can be included in the analysis as a reduced probability that units will be available.

The impact of distributed generation

The analysis of both the generation adequacy and the coordination issues is based upon the assumption that electricity is generated in large-scale facilities. A significant market penetration of distributed generation would alter both the generation adequacy and the coordination issues.

Distributed generation improves generation adequacy in three ways. First, the lead time for new generation units could be shortened dramatically if distributed generation gains a large market share. Presumably, permits would be standardized and therefore easy to obtain and a high volume of small generation units would allow mass production, delivery from storage and easy transport. As a result, markets could react much faster to shortages, reducing the tendency to investment cycles. Secondly, to the extent that a shift towards distributed generation means that there are more supply companies active in the market, it would limit the incentives for capacity withholding during periods of scarcity. The smaller the market share of a company, the less it is able to influence the market price. However, distributed generation does not necessarily mean that there are a large number of generating companies active in the market: there are economic reasons why the operation of these units would be aggregated by a limited number of companies who trade the electricity in the market (Kirby and Hirst, 2000). A third effect is that many small, distributed generators together have a higher reliability than the same amount of capacity in large-scale plants. Therefore, a smaller reserve margin would be needed to

obtain the same degree of reliability.

Distributed generation also has some complicating factors. Renewable energy sources like solar and wind energy increase the uncertainty of the availability of supply. If many consumers alternately take and inject electricity from and into the distribution network, metering and network operation will need to become much more sophisticated. Among others, real-time metering will be imperative in order to provide efficient incentives. The environmental impact of distributed generation might prove an obstacle: for instance, emissions in the case of natural gas, or the visual impact of wind turbines.

In a system of distributed generation, the question remains as to whether the total volume of generation capacity will be sufficient to meet rare demand peaks. The fundamental dynamics of the generation adequacy issue remain the same even in a fully decentralized system in which all electricity is produced with small units at the distribution network level. An exception would be if the units would be dimensioned on other criteria than peak electricity generation capacity, for instance, on heat production.

With respect to the issue of coordination, distributed generation could simplify matters. A large number of small generating units near consumers would reduce the demand for network capacity. The variations in load flows that are observed in current markets probably would be reduced for two reasons. First, output changes by the many small generation units would tend to cancel each other out. Second, the generating units would be closely linked to demand. Reactive power would be locally available although operational control could prove complicated with so many active units. These positive effects of distributed generation reinforce the argument made in Chapter 9 that network charges need to be more cost-reflective than they are currently in Europe.

Conclusions

Relaxation of the assumptions will not change the analysis substantially. The exception is the assumption that the role of distributed generation will be limited. In the long term, the validity of this assumption can certainly be questioned. However, penetration of distributed generation into current markets will likely take longer than it will take for an investment cycle to develop, so for the intermediate period the analysis of the generation adequacy issue holds.

11.5 The limits of competition

The analysis in this study has raised significant issues regarding liberalization. Separating a vertically integrated infrastructure sector into a competitive market and a regulated infrastructure turns out to be much less simple than was assumed at the outset of liberalization. Liberalization of electricity markets appears to stretch the limits of the market paradigm.

The delivery of electricity to consumers requires a combination of products and services only some of which can be provided competitively. There is room for discretion with respect to which functions are to be provided competitively and which ones are to be part

of the monopoly. Markets can be created for electric energy, generation capacity, reactive power, balancing power, interconnector capacity, retail, and metering services. The market for energy is closely related to fuel markets and, if present, emission credit markets. In addition, there are several monopoly functions, such as transmission, distribution and system balancing.

Electricity markets have the disadvantage that they require policy intervention to ensure generation adequacy. Markets for network-related services, such as auctions of congested interconnector capacity, are complicated by network externalities and limited by their transaction costs. The possibility of creating markets for reactive power and balancing power are limited by market power issues. The technical complexity of the electricity system is reflected in its economic organization by a web of related markets and monopoly functions.

Complexity

The more functions within the electricity system for which competition is introduced, the complexer the system becomes. Each sub-market requires rules, oversight and mechanisms in order to coordinate it with other functions in the system. For instance, if reactive power is provided competitively, there is the issue of local market power. An alternative would be to regulate the provision of reactive power or to let the TSO provide it himself. This might be more expensive but would make the market more transparent and might also result in better reliability of service. Another alternative is not to unbundle fully but to allow generators a financial stake in local networks so they have an interest in providing reactive power efficiently.

The tradeoff between achieving efficiency through the introduction of competition and the resulting complexity also exists at a more general level. In the end, the policy of liberalization electricity markets is paradoxical: it is based upon the assumption that government cannot regulate the electricity sector efficiently; if this is true, however, how can we expect government to regulate a liberalized market efficiently, which is very much more complex (Price C. Watts, 2001)? Electricity markets are highly complex and some questions, such as how to control market power and provide efficient incentives for transmission expansion, have not even been answered in theory. As the dynamics of liberalized electricity markets are not fully understood, the decentralization of control, which liberalization brings about, increases the risk of system failure. This is a reason to implement safety mechanisms such as a capacity mechanism, at least during the lengthy and difficult transition phase from a monopoly to a competitive market.

The complexity of electricity markets with full retail competition is perhaps a reason to reconsider a simpler model of liberalization. The greatest potential gains from competition arguably are found in the generation market. Therefore, the single buyer model – in which a competitive market in generation sells to regional monopolies who control all the other functions in the supply chain – combines simplicity with a large part of the potential efficiency gains from liberalization (cf. the ‘purchasing agency’ model in Hunt and Shuttleworth, 1996). Seen in this light, this unpopular model deserves second thought. Downsides are that the lack of consumer choice removes incentives for

improved consumer-friendliness in the supply of electricity. It would also conflict with the markets for 'green' electricity. This explains why this option currently has been removed as an option in the EU, while it also is not part of FERC's Standard Market Design (Directive 2003/54/EC; Fernandez, 2002).

Theoretical efficiency versus transparency

In the trade-off between a theoretically sound system and robustness, the USA tends to choose a theoretically sound approach, as exemplified by PJM's system of locational marginal pricing, while Europe has chosen a market model that is intended to be more simple and transparent. The advantage of locational marginal pricing is that it is a consistent system that should provide the correct incentives. Its disadvantage is that its complexity makes it intransparent. The algorithm used by the market operator to calculate all the locational marginal prices resembles the optimization programs that vertically integrated utilities use to calculate the optimal dispatch of their generation units. Consequently, locational marginal pricing is criticized for being too much of a centralist approach (Wu et al., 1996).

The European model of fixed transmission tariffs in combination with a market based upon bilateral contracts and voluntary spot markets leaves more room for private initiatives. However, in this model the network externalities are a significant obstacle to efficient long-term development. In the absence of a consistent system, other than locational marginal pricing, for addressing these externalities, the European model requires a combination of *ad hoc* corrections such as congestion management methods, variations in connection charges and perhaps permit requirements for locating generators or large loads that takes network effects into account. These solutions necessarily are rough approximations of the ideal incentives. Ironically, the more refined they become, such as coordinated auctions, the more complex and intransparent they become. In the end, they may start to resemble locational marginal pricing in their information requirements as well as in their complexity, while they may become even less transparent.

Differences in time constants

A final limit to the effectiveness of competition, or financial incentives in general, is created by the large time constants in the electricity sector. Markets generally have a substantially shorter time horizon than the life span of generation and network assets. It has been seen that in the case of generation this may give rise to investment cycles (see Chapter 5). These cycles may have a long duration and do not necessarily dampen over time. Even though the average investment level over the full cycle may be optimal, the periods of insufficient generation capacity may cause significant economic losses. The existence of optimal incentives alone is not sufficient to guarantee optimal behavior: the resulting system must also be stable.

11.6 Implications for other sectors

What lessons does this analysis provide for other economic sectors? The main theme of this research is that the technical characteristics of the system must be considered in the design of the market, lest a discrepancy develops between the market results and the technical possibilities. This may result in inefficient operating and investment decisions and in opportunities for strategic manipulation.

A second lesson concerns the liberalization of infrastructure sectors in which part of the chain of production remains a natural monopoly. The loss to coordination may be significant when infrastructures are unbundled. This loss may not be apparent in the beginning, as it may manifest itself mainly in the long term. However, it may offset the efficiency gains that were obtained from the introduction of competition. In principle, a decision to liberalize should therefore be founded upon a benefit-cost analysis of liberalization itself to determine whether the expected gains from liberalization outweigh the costs.

A third lesson is that a benefit-cost analysis of a liberalization policy should also include a risk analysis, which is an element that is too often absent from public policy. Not only should the expected benefits and costs be determined but also the risks involved with the change of the structure of the sector. Liberalization replaces hierarchical control with decentralized control, where the interest of the many actors should lead to a socially optimal outcome if they operate under efficient incentives. Liberalization changes the influence of public policy from determining output conditions to establishing process conditions. However, a sector like the electricity market is so complex that it is not possible to foresee all possible developments. Consequently, there is the risk that the market will not be designed optimally. As part of the decision to liberalize, the risks of implementing a flawed market design should be balanced against the expected benefits. From a social point of view, there may be a risk asymmetry with the costs of market failure far exceeding the costs of an inefficient monopoly. For the electricity sector, one might ask whether the promise of a modest average decrease in electricity price, better customer service and the possibility of choosing green electricity outweighs the increased risk of system failure.

Liberalization inevitably brings about regulatory uncertainty, which discourages investment. Care should be taken during the transmission phase so that the newly competitive market is not destabilized. In this respect, there is a dilemma between a ‘big bang’ approach to restructuring, in which a complete market design is imposed at once, and an evolutionary approach to the design of the newly liberalized market. A ‘big bang’ approach minimizes the period of regulatory uncertainty but it is difficult to make a good enough market design at once. Adjusting the market design as one goes along may lead to a better market design but may unacceptably prolong the period of regulatory uncertainty. (If done wrong, it may also lead to a compilation of *ad hoc* measures that results in an intransparent, inconsistent market design.)

A final observation is that connected markets should be liberalized at the same speed and with similar, if not the same, rules. Markets that are opened before their connected

markets expose themselves during the transition phase to strategic manipulation with permanent negative consequences. An example is the rapid foreign expansion of several large incumbent utility companies in Europe that, while they still are publicly owned, took over electricity firms in other countries, forming an oligopoly before competition has a chance to develop. A second problem is that differences in taxes and subsidies create artificial price differences between connected systems, leading to uneconomic flows.

12 Conclusions

This chapter summarizes the conclusions of this study. The first two sections present the conclusions regarding the issues of generation adequacy and coordination. Section 12.3 provides some more general conclusions with respect to the design of electricity generation markets. Section 12.4 provides suggestions for further research. The main conclusions are printed in ***bold*** *type; policy advice is indicated with an arrow:* ➢.

12.1 Generation adequacy

The California crisis

While there were many complicating and aggravating circumstances, the root cause of the electricity crisis in California was insufficient investment in generation capacity. Three other aspects stand out:

- The crisis was precipitated by the sudden reduction of exports to California from neighboring states (which had different market structures allowing them to give preference to their own consumers when electricity supply became tight).
- The crisis was severely aggravated by the strategic withholding of generation capacity, which increased the price substantially and caused a significant portion of the supply interruptions.
- The social costs of the crisis were not only a consequence of the interruptions of power supply but to a large extent due to the fact that the prices were far above their historical levels for about a year.
- The near complete absence of forward contracts made the retail companies vulnerable to high wholesale prices. The combination with fixed retail prices caused financial disaster among the retail companies, for which the tax payers paid the price.

These factors, except perhaps the last one, are not unique to California, which is reason for concern about generation adequacy in other energy-only markets.

Generation adequacy in energy-only markets

The low price-elasticity of demand and the inability to store electricity cause electricity prices in energy-only markets to be highly volatile. This causes significant investment

risk which is further increased in many cases by a lack of market transparency and regulatory uncertainty. Regulatory uncertainty is not only caused by the expectation that the rules of the electricity market and those of related markets (such as markets for fuels or emissions credits) may change but also by the threat of a price cap during a period of high prices.

Given these risks, investors prefer to err on the side of less capacity, so there appears to be a tendency towards too little investment. Once a shortage develops in an energy-only market, the resulting high prices provide a corrective signal. The long lead time for new generation capacity, however, means that the investment reaction is delayed by several years, during which time scarcity and high prices will continue to exist.

Competitive energy-only markets are susceptible to the development of investment cycles.

From the perspective of society, the loss of welfare from deviations from the optimal volume of generation capacity is strongly asymmetric. The social costs of insufficient generation capacity appear to be at least an order of magnitude higher than the costs of excess capacity.

Given uncertainty, the interest of consumers is to err on the side of too much rather than too little generation capacity.

Another disadvantage of relying upon volatile electricity prices for providing the investment signal is that these prices are susceptible to manipulation, especially during shortages. At these times, there is a strong incentive even for relatively small generating companies to increase the price by withholding generation capacity. This incentive is reduced by the presence of long-term contracts. However, neither the contract length or the degree to which they cover output appears sufficient to eliminate the incentive.

Price spikes, which should provide the investment signal in energy-only markets, are susceptible to manipulation.

When a capacity mechanism leads to a higher volume of generation capacity than is theoretically optimal, the frequency of price spikes and therefore also the development of generator market power would be reduced. The benefits of a more competitive market compensate at least partly for the cost of excess generation capacity. This is an additional reason for public policy to err on the side of excess generation capacity.

- **Electricity markets should have a capacity mechanism that stabilizes the volume of generation capacity and, consequently, electricity prices.**

The choice of capacity mechanism depends upon:
- whether it is to be implemented in an integrated or decentralized system,
- whether it is to be implemented in an open or closed market,
- how much time is available for its design, and
- whether consumers have or can get real-time meters.

Capacity mechanisms that directly influence the volume of generation capacity are more effective than those that use economic incentives because they are less vulnerable to information deficiencies. Of the capacity mechanisms that have been tried in practice, only PJM's system of capacity requirements works adequately. However, it does not provide an optimal incentive to generators to maximize output during a scarcity. Most important for European countries, capacity requirements do not appear effective in open, decentralized systems. The alternatives have not been tried in practice.

Decentralized, open markets require an innovative capacity mechanism, such as a variant of reliability contracts or financial capacity subscriptions.

Securing generation adequacy in an individual electricity system with significant exchanges is inevitably complicated. (See Section 8.2.) The capacity mechanism needs to be designed in such a way that the consumers who pay for the generation capacity also have access to this capacity when it is scarce. These consumers want to ensure that the generation capacity that was supported with the capacity mechanism is not used for exports during a regional shortage.

Joint implementation of a capacity mechanism by interconnected electricity systems would reduce, if not eliminate, the question of how to insulate the capacity mechanism against exchanges with neighboring systems. Joint implementation would also be economically more efficient. In this case, there are more options: for instance, if exchanges out of the system are relatively limited, capacity requirements would work in a decentralized system as well.

- **Clusters of strongly interconnected markets, such as Nordpool, the British isles, the Iberian peninsula, and the remaining bulk of the UCTE network, should implement a capacity mechanism jointly.**

- **The EU should take the initiative in implementing a capacity mechanism instead of the current policy of leaving it to subsidiarity.**

A deceptively attractive policy is to wait and see (monitor the market developments) and implement a capacity mechanism only when the need becomes clear. However, the long lead time for new generation capacity makes this a risky strategy. It means that the decision as to whether to implement a capacity mechanism depends upon projections of market developments for at least as many years in advance as it takes to implement the capacity mechanism and build new capacity. Moreover, implementation of a capacity mechanism in a market with a slim capacity margin requires additional transition measures.

- **Competitive energy-only markets should implement a capacity mechanism now, rather than wait until a lack of investment is observed.**

The analysis is based on the assumption of competitive behavior. In the case of an oligopolistic market – a common situation – generating companies may have a strategy of over-investment to deter new market entrants. This could lead to a sufficient level of

investment in generation capacity but it would also mean that the goal of liberalization, to increase economic efficiency through competitive pressure, was not achieved.

12.2 Coordination

To create a level playing field in the generation market, it is essential that generating companies are 'unbundled' from network companies in order to ensure that network companies have no economic interests in any generating company. Physically, however, there is a need to coordinate the operation and development of generation with the network. The issues are

- siting decision by generators,
- operational generator behavior, and
- the management of reactive power by generators.

As the system operator cannot directly influence operation of and investment in generation capacity, the ideal way to meet the need for coordination is to create efficient economic incentives for both the generators and the networks. One possible solution is locational marginal pricing.

Europe has chosen for a different solution, perhaps out of necessity: the institutional requirements for locational marginal pricing appear to be too high in the short term, considering the diverging legal, institutional and market structures of European electricity systems and the countries' reluctance to hand over control of their markets to a central authority. The European choice of fixed, separate transmission tariffs is intended to provide transparency and predictability to the market. However, in the absence of economically efficient incentives, a compilation of *ad hoc* measures will develop.

Paradoxically, the end result of *ex ante* fixed transmission tariffs may be not only less efficient than locational marginal pricing but also more complicated and less transparent.

The question of coordination is not only a consequence of imperfect incentives. The difference in time constants between the generation market and network further complicates matters. While the issue of generation adequacy is affected by the long lead time and life cycle of generation capacity, these are short in comparison to the long time for network investment and the life cycle of networks, in particular when the path-dependency of network design is considered. Therefore, the development of the generation stock needs to be coordinated with the network.

Generators need to receive locational incentives that stimulate efficient coordination with the network.

One of the consequences of *ex ante* fixed transmission tariffs is that the network may become congested. Congestion management methods are a means of allocating scarce network capacity. In doing so, they provide incentives to generation companies and to network operators. Unfortunately, none of the available methods provides efficient incentives to both sides.

Congestion pricing methods (essentially variations of auctions) are preferred to remedial methods such as redispatching.

Given the need to choose, it is more important to provide efficient incentives to generators than to the network managers. As the networks are regulated as natural monopolies, it is easier to guide their long-term development through means other than incentives. The European choice for explicit auctions is a good first step. The challenge will be how to refine this system so it makes better use of available network capacity without making the congestion management method overly complex.

12.3 General conclusions

The role of technology in market design

Generation adequacy and the coordination of the development of the network and the generating stock are both issues shaped by network externalities. This makes it difficult to create efficient economic incentives. These network externalities are a consequence of the technical characteristics of the electricity system. Other technical aspects, such as the need to balance supply, also play a role.

The technical characteristics of the electricity system must be considered in the design of the market.

Differences in time constants

The vastly different time frames within which the different parts of the system function present a challenge for system design. Electricity spot prices vary significantly by the hour; generation investment takes years and generators may last several decades; finally, network construction may have a lead time exceeding a decade, and networks have a path dependency in their development which causes the effects of design decisions to outlast the life cycles of the individual components, which themselves already may exceed half a century. Incentives that change rapidly, in comparison to the life cycle of the assets at hand, provide a risky basis for investment

Differences in time constants pose an obstacle to the effectiveness of financial incentives.

The system dynamics may be changed by technological developments. In particular, widespread introduction of distributed generation and the development of storage technologies could positively affect both adequacy and coordination.

Optimality versus robustness

The introduction of competition and other economic incentives is aimed at maximizing economic efficiency but includes a risk when these incentives do not function as intended. Hierarchical, direct control is more robust against unforeseen circumstances but less efficient. This trade-off occurs not only at the highest level – whether to liberalize or not – but also at other levels in the design of liberalized electricity markets, as there are

many functions that could, but not necessarily should, be provided competitively.

In highly complex sectors like the electricity sector, there is a trade-off between the economic efficiency and the relative simplicity of hierarchical control. This trade-off occurs at many levels in the institutional design of a sector.

Limits to the market

This research suggests that in the electricity sector, the limits of the liberalization paradigm have been encountered. Due to the complexity of the system and the network monopoly, the introduction of competition is accompanied by significant costs and risks. Liberalization policy appears to have been based upon steady-state estimates of potential economic gains without taking into consideration the increased risk of system failure caused by experimenting with a highly complex system. Complicating matters further is the fact that it is impossible to establish perfect market rules at the outset of liberalization, while adjusting the rules along the way creates regulatory uncertainty which undermines long-term system development.

- **Policy analysis should not only consider the expected benefits of restructuring a sector but also the risks and potential consequences of policy failure.**

12.4 Further research

Capacity mechanisms

If a capacity mechanism is to be implemented in an open, decentralized system, or if a more efficient mechanism than capacity requirements is desired, a new capacity mechanism needs to be developed. The best candidates are centralized and bilateral reliability contracts. However, these capacity mechanisms need to be designed in more detail before they can be implemented. Specifically, their dynamic behavior and their susceptibility to manipulation should be evaluated.

Coordination

The analysis of the coordination issue was limited to a framing of the issue. Much more research is to needed to understand the scale and severity of any of the described externalities in practice. Detailed modeling of specific networks is the only way to arrive at quantitative conclusions, which is perhaps why so little research has been done on the subject. Some of the questions to be addressed are the following:

- How do differences in market structures, taxes, subsidies, cross-subsidies and transmission tariffs in connected electricity systems affect inter-system trade?
- What is the cost of reactive power management in a liberalized market? To what extent is the operational security of the network impacted by the network managers' dependency upon independent, commercial generating companies?
- To what extent do the locational decisions of generators deviate from the most efficient scenario (both in terms of cost and system reliability)?

If a need to improve coordination between the generation market and the networks is identified, various policy instruments need to be studied. As these instruments would have a pragmatic nature, the only way to determine their effectiveness – apart from trying them in practice – would be through extensive and detailed modeling of the specific market in which they would be introduced.

Congestion management

With respect to congestion management, the European choice for explicit auctions appears a reasonable first solution. Current research focuses on refining the auctions but this entails a risk of still not having a high enough geographic resolution, while becoming complex and intransparent. A drawback of any kind of explicit auction remains the need for separate network and energy transactions, which is a significant barrier to trade. Therefore the feasibility and merits of systems that combine both, such as market splitting and locational marginal pricing, should be explored.

References

Aalbers, R.F.T., Bressers, D.L.F., Dijkgraaf, E., Hoogendoorn P.J. and De Klerk S.C. 1999. *Een level playing field op de Nederlandse elektriciteitsmarkt, een tariefstructuur voor het netgebruik*. Rotterdam, OCFEB Research Centre for Economic Policy.

Abbott, M. 2001. 'Is the Security of Electricity Supply a Public Good?'. *The Electricity Journal* 14 (7): 31-33.

Ackermann, T., Andersson, G. and Söder, L. 2001. 'Distributed generation: a definition'. *Electric Power Systems Research* 57: 195-204.

AER (Algemene Energieraad) 2003. *Energiemarkten op de weegschaal, signaleringsadvies van de energieraad over de liberalisering van de Europese elektriciteitsmarkt*. The Hague: AER.

Ajodhia, V., Hakvoort, R.A. and Van Gemert, M. 2002. 'Electricity Outage Cost Valuation: A Survey'. In: *Proceedings of CEPSI 2002*, Fukuoka, Japan.

Ajodhia, V. 2002a. 'Regulating Electricity Networks: Yardstick Competition and Reliability of Supply'. In: *Proceedings, 22nd USAEE/IAEE North American Conference,* 6-8 October, Vancouver, B.C.

Ajodhia, V. 2002b. 'Integrated Price and Reliability Regulation: The European Experience'. In: *Proceedings, IEEE /PES T and D 2002 Asia Pacific*, Yokohama.

Allaz, B. and Vila, J.-L. 1993. 'Cournot Competition, Forward Markets and Efficiency'. *Journal of Economic Theory* 59 (1): 1-16.

Allen, M. and Booth, W. 2001. 'Spread of Calif. Crisis Concerns Bush, Western Governors Get Assurances of Action'. *Washington Post,* January 30.

Audouin, R., Chaniotis, D., Tsamasphyrou, P. and Coulondre, J-M. 2002. 'Coordinated auctioning of cross-border capacity: an implementation'. In: *Proceedings of the Fifth International Conference on Power System Management and Control (PSMC)*, London, IEE,: 25-30.

Australian Competition and Consumer Commission 2000. *Determination, Applications for Authorisation; VoLL, Capacity Mechanisms and Price Floor*. File nr. C1999/865. Obtained from: www.accc.gov.au/electric/authorisations/previous%5Fdeter/voll_pricing_floor_capacity_mechanisms/VoLL_CM_pf.htm.

Averch, H. and Johnson, L.L. 1962. 'Behavior of the firm under regulatory constraint'. *The American Economic Review* 52: 1053-1069.

Behr, P. 2001. 'AES Outage in California Probed'. *Washington Post*, March 16.

Behr, P. 2002. 'Papers Show That Enron Manipulated Calif. Crisis'. *Washington Post*, May 7.

Berry, J.M. 2001. 'U.S. Officials: Impact of Calif.'s Crisis Muted for Now'. *Washington Post,* January 20.

Besser, J.G., Farr, J.G. and Tierney, S.F. 2002. 'The Political Economy of Long-Term Generation Adequacy: Why an ICAP Mechanism is Needed as Part of Standard Market Design'. *The Electricity Journal* 15 (7): 53-62.

Bidwell, M. and Henney, A. 2003. *Long-term generation adequacy through reliability options.* presentation at the Dutch Ministry of Economic Affairs, The Hague, October 10.

Bijvoet, C., De Nooij, M. and Koopmans, C. 2003. *Gansch het raderwerk staat stil. De kosten van stroomstoringen*. Amsterdam, Stichting Economisch Onderzoek (SEO). Obtained from: www.tennet.nl/images/14_5477.pdf.

Billinton, R. 1994. 'Evaluation of reliability worth in an electric power system'. *Reliability Engineering and System Safety*. 46: 15-23.

Billinton, R. and Allan, R.N. 1984. *Reliability Evaluation of Power Systems*. London, Pitman Publishing Limited.

Billinton, R. and Allan, R.N. 1992. *Reliability Evaluation of Engineering Systems, Concepts and Techniques*. New York, Plenum Press.

Billinton, R., Allan, R.N. and Salvaderi, L. (eds.) 1991. *Applied Reliability Assessment in Electric Power Systems*. New York, IEEE.

Bjørndal, M. and Jørnsten, K. 2000. **'**Investment Paradoxes in Electricity Networks'. In: *Proceedings, IAEE European Conference 2000: Towards an Integrated European Energy Market*.

Bonneville Power Administration 2001. *Cold weather spurs energy consumption; more power needed from Columbia River dams*. News release, February 13.

Borenstein, S. 2001. 'The trouble with electricity markets (and some solutions)'. *POWER Working Paper PWP-081*, University of California at Berkeley. Obtained from: www.ucei.berkeley.edu/ucei/pubs-pwp.html.

Borenstein, S., Bushnell, J. and Stoft, S. 2000. 'The competitive effects of transmission capacity in a deregulated electricity industry'. *RAND Journal of Economics*, 31 (2): 294-325.

Borenstein, S. and Holland, S.P. 2002. *Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Prices*. University of California at Berkeley, Center for the Study of Energy Markets.

Botterud, A., Korpås, M., Vogstad, K.-O. and Wangensteen, I. 2002. 'A Dynamic Simulation Model for Long-Term Analysis of the Power Market'. In: *Proceedings, 14th Power Systems*

Computation Conference (PSCC '02), Sevilla.

Bowring, J.E. and Gramlich, R.E. 2000. 'The Role of Capacity Obligations in a Restructured Pennsylvania-New Jersey-Maryland Electricity'. *The Electricity Journal* 13 (9): 57-67.

Budhraja, V. 2003. 'Harmonizing Electricity Markets with the Physics of Electricity', *The Electricity Journal* 16 (3): 51-58.

The Brattle Group 2003. *The Potential for a Dutch Operating Reserves Market.* Report to the Office of Energy Regulation (DTe), London, The Brattle Group.

Bushnell, J. 2003. *California's Electricity Crisis: A Market Apart?* Berkeley, University of California Energy Institute, Center for the Study of Energy Markets Working Paper 119. Obtained from: www.ucei.berkeley.edu/PDF/csemwp119.pdf.

CAISO (California Independent System Operator) 2000. *Report on California Energy Market Issues and Performance: May-June,2000,* California Independent System Operator, Department of Market Analysis. Obtained from: www.caiso.com/docs/09003a6080/07/40/09003a6080074029.pdf.

CAISO (California Independent System Operator) 2001. *CAISO Summer 2001 Assessment*. Folsom (California) CAISO.
Obtained from: www.caiso.com/docs/09003a6080/0c/af/09003a60800cafcd.pdf.

California State Senate 2002. Web site of the committee to Investigate Price Manipulation of the Wholesale Energy Market:
http://www.sen.ca.gov/ftp/sen/committee/select/INVESTIGATE/_home1/PROFILE.HTM.

Camfield, R.J. and Schuster, A.G. 2000. 'Pricing Transmission Services Efficiently'. *The Electricity Journal* 13 (9): 13-32.

Caramanis, M.C. 1982. 'Investment decisions and long-term planning under electricity spot pricing'. *IEEE Transactions on Power Apparatus and Systems* 101 (12): 4640-4648.

Caramanis, M.C., Bohn, R.E. and Schweppe, F.C. 1982. 'Optimal Spot Pricing: Practice and Theory'. *IEEE Transactions on Power Apparatus and Systems PAS-101* (9): 3234-3245.

Carere, E., Fox-Penner, P., Lapuerta C. and Moselle, B. 2001. *The California Crisis and its Lessons for the EU.* London, The Brattle Group.

Castro-Rodriguez, F., Marín Uribe, P. and Siotis, G. 2001. *Capacity Choices in Liberalized Electricity Markets*. CEPR Discussion Paper no. 2998, London, Centre for Economic Policy Research. Obtained from: www.cepr.org/pubs/dps/DP2998.asp.

Cazalet, E.G., Clark, C.E. and Keelin, T.W. 1978. *Costs and Benefits of Over/Under Capacity in Electric Power System Planning*. Palo Alto (California), EPRI.

CBO (Congress of the United States, Congressional Budget Office) 2001. *Causes and Lessons of the California Electricity Crisis*. Washington, DC, Congress of the United States, Congressional Budget Office.

CEC (California Energy Commission) 1998. *New Options for Agricultural Customers: California's Electric Industry Restructuring*. State of California Energy Commission P400-97-005.

CEC (California Energy Commission) 2000. *California Energy Demand 2000 – 2010, Technical Report to California Energy Outlook*. Docket #99-CEO-1. Obtained from: www.energy.ca.gov/reports/2000-07-14_200-00-002.PDF.

CEC (California Energy Commission) 2001a. Web site: www.energy.ca.gov, data on new generation projects: www.energy.ca.gov/sitingcases/approved.html#chart1.

CEC (California Energy Commission) 2001b. Energy Facilities Siting/Licensing Process web site: www.energy.ca.gov/sitingcases/index.html.

Chao, X.Y., Feng, X.M. and Slump, D.J. 1999. 'Impact of Deregulation on Power Delivery Planning'. *1999 IEEE Transmission and Distribution Conference Proceedings*, *Vol. 1*: 340-344.

Chao, H.P., Peck, S., Oren, S.S. and Wilson, R. 2000. 'Flow-gate Transmission Rights and Congestion Management'. *The Electricity Journal* 13 (8): 38-58.

Coleman, J. 2001. 'California Conservation Working'. *Associated Press*, August 17.

Commissie CO_2-handel 2002. *Handelen voor een beter milieu, Haalbaarheid van CO2-emissiehandel in Nederland*, Report to the Minister of VROM (public housing, spatial planning and the environment). Obtained from: www.co2handel.nl/docs/eindadvies.pdf.

CPUC (California Public Utilities Commission) 2002. *Report on Wholesale Electric Generation Investigation*. San Francisco, CPUC.

Crampes, C. and Laffont, J.-J. 2001. 'Transport pricing in the electricity industry'. *Oxford Review of Economic Policy* 17 (3): 313-328.

Crew, M.A. and Kleindorfer, P.R. 1985. 'Governance Structures for Natural Monopoly'. *Journal of Behavioral Economics*, 14 (0): 117-140.

Day, C.J., Hobbs, B.F. and Pang, J.S. 2002. 'Oligopolistic Competition in Power Networks: A Conjectured Supply Function Approach'. *IEEE Trans. Power Systems* 17 (3): 597-607.

De Vries, L.J. 2001. 'Long-term investment in electricity networks: mapping the issues'. In: *Proceedings, Critical Infrastructures Conference*, The Hague, June 27 – 29.

De Vries, L.J. 2003. 'Infrastructure Investment after Liberalization'. In: *Thissen, W.A.H., Herder, P.M. (Eds.) Critical infrastructures - State of the art in research and application*, Dordrecht, Kluwer Academic Publishers: 163-179.

De Vries, L.J. 2004. 'Policy Framework for the Stabilization of Investment in Generating Capacity'. In: *Proceedings, 19th World Energy Congress and Exhibition*, Sydney, 5-9 September 2004.

De Vries, L.J. and Hakvoort, R.A. 2002a. 'Market failure in generation investment? The Dutch perspective'. In: *Proceedings of the Fifth International Conference on Power System Management and Control (PSMC),* London, IEE, 17 - 19 April: 7-12.

De Vries, L.J. and Hakvoort, R.A. 2002b. 'An Economic Assessment of Congestion Management Methods for Electricity Transmission Networks'. *Journal of Network Industries* 3 (4): 425-466.

De Vries, L.J. and Hakvoort, R.A. 2003a. 'Opties voor voorzieningszekerheid'. *Economisch Statistische Berichten (ESB)* 87 (4396; 7 March): 108-111.

De Vries, L.J. and Hakvoort, R.A. 2003b. 'The question of generation adequacy in liberalized electricity markets'. In: *Proceedings, 26th IAEE International Conference,* Prague, 4-7 June.

De Vries, L.J. and Hakvoort, R.A. 2003c. 'Generation adequacy in Europe: a policy framework'. In: *Proceedings of PowerCon 2003 "Blackout" Conference (IASTED),* 10-12 December, New York, 114-119.

De Vries, L.J., Knops, H.P.A. and Hakvoort, R.A. 2004. Bilateral Reliability Contracts: An Innovative Approach to Maintaining Generation Adequacy in Liberalized Electricity Markets. In: *Proceedings, IRAEE Conference 'Energy and Security in the Changing World'*, Tehran, May 25-27.

Dixit, A.K. and Pindyck, R.S. 1994. *Investment under Uncertainty*. Princeton, New Jersey, Princeton University Press.

Directive 96/92/EC of the European Parliament and of the Council of 19 December 1996 concerning common rules for the internal market in electricity. *Official Journal of the European Union,* 1997, L 27: 20-29.

Directive 2003/54/EC of the European Parliament and of the Council of 26 June 2003 concerning common rules for the internal market in electricity and repealing Directive 96/92/EC. *Official Journal of the European Union*, 2003. L 176: 37-55.

DTe 2001. *Advies van de DTe aan de Minister van Economische Zaken inzake de leveringszekerheid van de Nederlandse elektriciteitsvoorziening op de lange termijn ('Advice from the Dutch Energy Regulator to the Minister of Economic Affairs regarding the security of the Dutch electricity supply in the long term')*. The Hague, DTe (Office of Energy Regulation).

Dondi, P., Bayoumi, D., Haederli, C., Julian, D. and Suter, M. 2002. 'Network integration of distributed power generation'. *Journal of Power Sources* 106: 1-9.

Doorman, G. 2000. *Peaking Capacity in Restructured Power Systems*. Thesis (Ph.D.), Norwegian University of Science and Technology, Faculty of Electrical Engineering and Telecommunications, Department of Electrical Power Engineering.

EC (Commission of the European Communities) 1999. *Second Report to the Council and the European Parliament on Harmonisation Requirements (Directive 96/92/EC concerning common rules for the internal market in electricity)*. Brussels, EC.

EC (Commission of the European Communities) 2000. *Conclusions, Sixth meeting of the European electricity Regulatory Forum*, Florence, EC.

EC (Commission of the European Communities) 2001a. *Commission Staff Working Paper: Completing the internal energy market*. Brussels, EC (SEC (2001) 438), obtained from: http://europa.eu.int/comm/energy/library/438.pdf.

EC (Commission of the European Communities) 2001b. *Communication from the Commission to the Council and the European Parliament: European Energy Infrastructure, COM(2001) 775 final, 2001/0311(COD)*. Brussels, EC.

EC (Commission of the European Communities) 2001c. *Proposal for a Regulation of the European Parliament and of the Council on conditions for access to the network for cross-border exchanges in electricity*. Brussels, EC. Obtained from: http://europa.eu.int/comm/energy/en/internal-market/int-market.html.

EC (Commission of the European Communities) 2001d. *Proposal for a Regulation of the European Parliament and of the Council on conditions for access to the network for cross-border exchanges in electricity, Explanatory Memorandum*. Brussels, EC. Obtained from: http://europa.eu.int/comm/energy/en/internal-market/library/reglement_en_acte.pdf.

EC (Commission of the European Communities, DG TREN) 2002a. *Congestion Management in the EU Electricity Transmission Network – Status Report (September 2002)***,** obtained from: http://europa.eu.int/comm/energy/en/elec_single_market/florence9/discussion_paper/congestion_management.pdf.

EC (Commission of the European Communities) 2002b. *Second benchmarking report on the implementation of the internal electricity and gas market*, *Commission Staff Working* Paper. Brussels, EC. Obtained from: http://europa.eu.int/comm/energy/en/gas_single_market/2benchmarking/sec_2002_1038_en.pdf.

EC (Commission of the European Communities) 2002c. *Amended proposal for a Directive of the European Parliament and of the Council establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC, COM(2002) 680 final*. Brussels, EC. Obtained from: http://europa.eu.int/eur-lex/en/com/pdf/2002/com2002_0680en01.pdf.

EIA (Energy Information Administration) 2002. *Status of the California Electricity Situation*. Washington, DC, US Department of Energy, Energy Information Administration. Obtained from: www.eia.doe.gov/cneaf/electricity/california/california.html.

EnergieManagement 2003. *NMa: Nuon moet 900 MW productiecapaciteit veilen*. Obtained from: www.energiemanagement.nl, November 28.

EnergieNed 2002. *De energievoorziening in goede handen, eerste bevindingen liberalisering energiemarkt*. Arnhem, EnergieNed.

EnergieNed 2003. *Condities voor een betrouwbare energievoorziening, eerste bevindingen waarborgen voorzieningszekerheid*. Arnhem, EnergieNed.

EnergieNed 2004. *Energy in the Netherlands 2003, Facts and Figures*. Arnhem, EnergieNed.

EPRI 2001. *The Western States Power Crisis: Imperatives and Opportunities*. Palo Alto (California), EPRI.

ETSO (European Transmission System Operators) 1999. *Evaluation of congestion management methods for cross-border transmission*. Brussels, ETSO. Obtained from: www.etso-net.org.

ETSO (European Transmission System Operators) 2001a. *Definitions of Transfer Capacities in Liberalized Electricity Markets*. Brussels, ETSO. Obtained from: www.etso-net.org/media/download/Transfer%20Capacity%20Definitions.pdf.

ETSO (European Transmission System Operators) 2001b. *Co-ordinated Auctioning, a market-*

based method for transmission capacity allocation in meshed networks. Brussels, ETSO. Obtained from: www.etso-net.org/media/download/Coordinated%20Auctioning.pdf.

'EU energy markets face cohesion barriers'. *Power Europe* (4), 4 January 2002: 5.

European Council 2002. *Presidency Conclusions.* Barcelona, 15-16 March.

FERC 2000. *Staff Report to the Federal energy Regulatory Commission on Western Markets and the Causes of the Summer 2000 Price Abnormalities*. Washington, DC, FERC. Obtained from: www.stoft.com/x/cal/20001101-FERC-staff-all.pdf.

FERC 2002a. *Initial Report on Company-Specific Separate Proceedings and Generic Reevaluations; Published Natural Gas Price Data; and Enron Trading Strategies; Fact-Finding Investigation of Potential Manipulation of Electric and Natural Gas Prices*. Washington, DC, FERC, Docket No. PA02-2-000. Obtained from: www.ferc.fed.us/electric/bulkpower/pa02-2/Initial-Report-PA02-2-000.pdf.

FERC 2002b. *Remedying Undue Discrimination through Open Access Transmission Service and Standard Electricity Market Design, Notice of Proposed Rulemaking*. Washington, DC, FERC, Docket No. RM01-12-000. Obtained from: www.ferc.gov/Electric/RTO/Mrkt-Strct-comments/nopr/Web-NOPR.pdf.

FERC, 2002c. *Public Utilities Commission of the State of California vs. El Paso, Initial Decision*. Washington, DC, FERC. Obtained from: www.ferc.gov/RP00-241-006-09-23-02.pdf.

Fernandez, A. 2002. *An Overview of FERC's Standard Market Design NOPR* (power point presentation). Washington, DC, FERC. Obtained from: www.ferc.gov/Electric/RTO/Mrkt-Strct-comments/NOPR/SMD-08-19-02.pdf.

Ford, A. 1999. 'Cycles in competitive electricity markets: a simulation study of the western United States'. *Energy Policy* (27): 637-658.

Ford, A. 2001. 'Waiting for the boom: a simulation study of power plant construction in California'. *Energy Policy* 29: 847-869.

Fraser, H. and Lo Passo, F. 2003. 'Developing a Capacity Payment Mechanism in Italy', *The Electricity Journal,* 16 (9): 54-58.

Gladstone, M. and Bailey, B. 2000. 'State's Long Road to Current Problems'. *San Jose Mercury News*, November 30. Obtained from: www0.mercurycenter.com/.

Goel, L. and Billinton, R. 1997. 'Impacts of pertinent factors on reliability worth indices in an electric power system'. *Electric Power Systems Research* 41: 151-158.

Hakvoort, R.A. 2000. 'Liberalisation of the Power Sector: What Does It Really Mean?'. In: *13th Annual Western Conference 'Competitive challenge in Network Industries'*, Monterrey, California, 5-7 July.

Hakvoort, R.A. and De Vries, L.J. 2002. 'Opportunities and Threats for Electricity Network Companies in a Restructured Power Market'. In: *Proceedings of the 14th Conference of the Electric Power Supply Industry (CEPSI),* Fukuoka.

Harvey, S.M., Hogan, W.W. and Pope, S.L. 1996. *Transmission Capacity Reservations and Transmission Congestion Contracts.* Mimeo, Cambridge, Massachussetts, Harvard University. Obtained from: http://www.whogan.com/.

Haubrich, H.J., Fritz, W. and Vennegeerts, H. 1999. *Study on Cross-border electricity transmission tariffs by order of the European Commission, DG XVII/C1*. University of Aachen.

Hawkins, D., *The California Report*. PowerPoint presentation, California ISO, Oct. 2, 2001.

Hebert, H.J. 2001. 'Solution eludes power players, State and federal officials, utilities representatives and power producers and brokers hold a meeting to try to ease California's electricity crisis'. *The Associated Press*, January 10.

Helm, D.R. 2001. 'The Assessment: European Networks – Competition, Interconnection, and Regulation'. *Oxford Review of Economic Policy* 17 (3): 297-312.

Henney, A. 2004. 'Will NETA ensure generation adequacy', *Power UK* 122: 10-26.

Hesselmans, A.N. 1995. *'De ware ingenieur'. Clarence Feldmann, Delfts hoogleraar en grondlegger van de provinciale elektriciteitsvoorziening*, Utrecht, Stichting Histosearch.

Hirst, E. 2000. 'Do We Need More Transmission Capacity?'. *The Electricity Journal* 13 (9): 78-89.

Hirst, E. 2001. *The California electricity crisis: lessons for other states*, Oak Ridge (Tennessee), Consulting in Electric-Industry Restructuring, July 2001.

Hirst, E. and Hadley, S. 1999. 'Generation Adequacy: Who Decides?'. *The Electricity Journal* 12 (8): 11-21.

Hirst, E. and Kirby, B. 1997. *Creating Competitive Markets for Ancillary Services*. Report prepared for the U.S. Department of Energy, Oak Ridge, Tennessee, Oak Ridge National Laboratory.

Hirst, E. and Kirby, B. 2001. *Retail-Load Participation in Competitive Wholesale Electricity Markets*, Report Prepared for the Edison Electric Institute, Washington, D.C. and the Project for Sustainable FERC Energy Policy, Alexandria, Virginia.

Hobbs, B., Iñón, J. and Kahal, M. 2001a. 'Issues concerning ICAP and alternative approaches for power capacity markets'. In: *Proceedings of the Market Design 2001 Conference, Stockholm 7 and 8 June 2001*: 7 – 18.

Hobbs, B., Iñón, J. and Stoft, S.E. 2001b. 'Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?'. *The Electricity Journal* 14 (6): 23-34.

Hobbs, B.F., Iñón, J. and Kahal, M. 2001c. *A Review of Issues Concerning Electric Power Capacity Markets*, *Project report submitted to the Maryland Power Plant Research Program, Maryland Department of Natural Resources*. Baltimore, The Johns Hopkins University.

Hobbs, B.F., Rijkers, F.A.M. and Wals, A.F. 2004. 'Strategic Generation with Conjectured Transmission Price Responses in a Mixed Transmission Pricing System II: Application', *IEEE Transactions on Power Systems (forthcoming)*. Obtained from: www.jhu.edu/~dogee/people/faculty/hobbs/IEEE_ECN_Part2_Aug03.pdf.

Hogan, W.W. 1992. 'Contract Networks for Electric Power Transmission'. *Journal of Regulatory Economics*, 4: 211-242.

Hogan, W.W. 1993. 'Market in Real Electric Networks Require Reactive Prices'. *The Energy Journal* 14 (3): 171-200.

Holson, L.M. and Oppel Jr., R.A. 2001. 'Trying to Follow the Money in California's Energy Mess'. *The New York Times*, January 12.

Hunt, S. 2002. *Making competition work in electricity*. New York, John Wiley & Sons, Inc.

Hunt, S. and Shuttleworth, G. 1996. *Competition and Choice in Electricity*, Chichester, John Wiley & Sons, Inc.

Jaffe, A.B., and Felder, F.A. 1996. 'Should Electricity Markets Have a Capacity Requirement? If So, How Should It Be Priced?'. *The Electricity Journal* 9 (10): 52-60.

Johnson, S. and Woolfolk, J. 2000. 'Power shortage worsens, state pushed to brink of black-outs as supply falls'. *San Jose Mercury News*, December 7. Obtained from: www.bayarea.com/mld/bayarea/archives//.

Johnson, S. and Woolfolk, J. 2001. 'Energy crisis turns the 'golden state' into a 'moneypit''. *San Jose Mercury News*, January 18. Obtained from: www.bayarea.com/mld/bayarea/archives//.

Jonnavithula, S. and Billinton, R. 1998. 'Cost-benefit analysis of generation additions in system planning'. In: *IEE Proceedings on Generation, Transmission and Distribution* 145 (3): 288-292.

Joskow, P. and Kahn, E. 2002. 'A Quantitative Analysis of Pricing Behavior in California's Wholesale Market During Summer 2000'. *The Energy Journal* 23 (4): 1-35. Obtained from: econ-www.mit.edu/faculty/pjoskow/files/JK_PaperREVISED.pdf.

Joskow, P.L. and Tirole, J. 2003. *Merchant Transmission Investment*. Working Paper 9534, Cambridge, MA, National Bureau of Economic Research. Obtained from: http://papers.nber.org/papers/w9534.pdf.

Kahn, J.R. 1998. *The Economic Approach to Environmental and Natural Resources*. 2nd ed., Fort Worth, The Dryden Press.

Kahn, E. and Baldick, R. 1994. 'Reactive Power is a Cheap Constraint', *The Energy Journal* 15 (4): 191-201.

Kahn, M. and Lynch, L. 2000. *California's electricity options and challenges, report to Governor Gray Davis*. Electricity Oversight Board and California Public Utilities Commission. Obtained from: www.cpuc.ca.gov/published/report/GOV_REPORT.htm.

Kaplan, T. and Guido, M. 2001. 'Blackouts roll across the Bay Area'. *San Jose Mercury News*, January 17. Obtained from: www.bayarea.com/mld/bayarea/archives//.

Kariuki, K.K. and Allan, R.N. 1996a. 'Evaluation of reliability worth and value of lost load'. *IEE Proceedings on Generation, Transmission and Distribution* 143 (2): 171-180.
Kariuki, K.K. and Allan, R.N. 1996b. 'Factors affecting customer outage costs due to electric service interruptions'. *IEE Proceedings on Generation, Transmission, Distribution* 143 (6): 521-

528.

Kirby, B. and Hirst, E. 2000. *Bulk-power reliability and commercial implications of distributed resources*. Oak Ridge, Tennessee, Oak Ridge National Laboratory. Obtained from: www.ornl.gov/ORNL/BTC/Restructuring/pub.htm.

Kirsch, L.D. and Singh, H. 1995. 'Pricing Ancillary Electric Power Services'. *The Electricity Journal* 8 (8): 28-36.

Kling, W.L. 1998. *Planning en Bedrijfsvoering van Elektriciteitsvoorzieningsystemen*. Delft University of Technology.

Knops, H.P.A., De Vries, L.J. and Hakvoort, R.A. 2001. 'Congestion management in the European electricity system: an evaluation of the alternatives'. *Journal of Network Industries* 2 (3-4): 311-351.

Knops, H.P.A. 2003. 'Weighing Ways of Keeping the Energy Balance'. In: *Proceedings, 26th Annual IAEE Conference*, Prague, June 4-7.

Kolstad, J. and Wolak, F. 2003. *Using Environmental Emissions Permit Prices to Raise Electricity Prices: Evidence from the California Electricity Market.* University of California Energy Institute, Center for the Study of Energy Markets, Working Paper 113. Obtained from: www.ucei.berkeley.edu/PDF/csemwp113.pdf.

Künneke, R.W. 1999. 'Electricity networks: how 'natural' is the monopoly?'. *Utilities Policy* 8: 99-108.

Künneke, R.W., Bouwmans, I., Kling, W.L., Van Poelje, H., Slootweg, J.G., Stout, H.D., De Vries, L.J., and Wolters, M. 2001. *Innovatie in energienetwerken*. Study commissioned by EnergieNed, Delft University of Technology.

Liedtke, M. 2000. 'State Regulators Cut San Diego Power Rates 3-2 Vote: Utilities Commission Approves Cap On Costs In Effort To Ease Shock Of Deregulation'. *Associated Press*, Aug. 22.

Lindqvist, C 2001. 'Methods to secure peak load capacity'. In: *Proceedings of the Market Design 2001 Conference,* Stockholm 7 - 8 June: 41-45.

Manifesto on the California Electricity Crisis, Generated and endorsed by an ad-hoc group of concerned professors, former public officials, and consultants, Convened under the auspices of the Institute of Management, Innovation, and Organization at the University of California, Berkeley 2001. Obtained from: haas.berkeley.edu/news/california_electricity_crisis.html.

Marshall, M. and McAllister, S. 2000. 'Potential outages could produce shocking costs, Tech firms alarmed by power crunch'. *San Jose Mercury News*, Dec. 9. Obtained from: www0.mercurycenter.com/.

Moore, P. and Ashmole, P. 1995. 'Flexible AC transmission systems'. *Power Engineering Journal,* December: 282-286.

Moore, P. and Ashmole, P. 1996. 'Flexible AC transmission systems, Part 2: Methods of transmission line compensation'. *Power Engineering Journal,* December: 273-278.

Moore, P. and Ashmole, P. 1997. 'Flexible AC transmission systems, Part 3: Conventional FACTS controllers'. *Power Engineering Journal* August: 177-183.

Moore, P. and Ashmole, P. 1998. 'Flexible AC Transmission Systems, Part 4: Advanced FACTS controllers'. *Power Engineering Journal,* April: 95-100.

Morgan, M.G. and Henrion, M. 1990. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*, Cambridge, Cambridge University Press.

Nasser, T.O. 1998. *Congestion Pricing and Network Expansion*. Policy Research Working Paper 1896, Washington, DC, The World Bank.

North American Electric Reliability Council 2004. Web site: www.nerc.com. Generating Availability Data System information: www.nerc.com/~gads.

Neuhoff, K. and De Vries, L.J. 2004. 'Insufficient Incentives for Investment in Electricity Generation'. Submitted to *Utilities Policy*.

New York ISO 2002. Web site: www.nyiso.com; section on ICAP: www.nyiso.com/markets/icapinfo.html.

Newbery, D.M. 2001. *Regulating electricity to ensure efficient competition*. Paper presented at the CEPR/ESRC Workshop on The Political Economy of Regulation, London, 1 November.

Newbery, D.M. 2002a. *Regulatory Challenges to European Electricity Liberalisation*. Paper presented at the conference Regulatory Reform - Remaining Challenges for Policy Makers, Stockholm, June 10.

Newbery, D.M. 2002b. Comments made at a workshop on security of supply, The Hague, Ministry of Economic Affairs, September 17.

Newbery, D., Van Damme, E. and Von der Fehr, N.-H.M. 2003. *MSC Analysis of effects of gas costs for electricity generation (unpublished)*, The Hague, DTe (Office of Energy Regulation), Market Surveillance Committee.

Nilssen, G. and Walther, B. 2001. 'Market-based Power Reserves Acquirement, An approach implemented in the Norwegian power system, with participation from both generators and large consumers'. In: *Proceedings of the Market Design 2001 Conference*, Stockholm, 7 - 8 June 2001: 59-66.

Nissenbaum, D., Devall, C. and Woolfolk, J. 2001. 'Gov. Davis announces 40 long-term contracts, Conservation still needed to prevent summer blackouts'. *San Jose Mercury News*, March 5. Obtained from: www0.mercurycenter.com/.

Nissenbaum, D. 2001. 'Power bills drain budget surplus'. *San Jose Mercury News*, May 8. Obtained from: www0.mercurycenter.com/.

NordPool web site: www.nordpool.com/products/elspot/index.html.

Ocaña, C. and Hariton, A. 2002. *Security of Supply in Electricity Markets, Evidence and Policy Issues*. Paris, IEA Publications.

O'Donnell, L. 2001. 'Practical Demand-side Flexibility to Save Peak Capacity, Provide Spinning Reserve and Respond to Capacity Shortfalls'. In: *Proceedings of the Market Design 2001 Conference*, Stockholm 7 - 8 June: 85-90.

Oppel, R.A. and Bergman, L. 2002. 'Judge Concludes Energy Company Drove Up Prices'. *The New York Times*, September 23.

Oren, S.S. 1998. 'Transmission Pricing and Congestion Management: Efficiency, Simplicity and Open Access'. In: *Proceedings of the EPRI Conference on Innovative Pricing, Washington DC, (June 17-19, 1998).*

Oren, S.S. 2000. 'Capacity Payments and Supply Adequacy in Competitive Electricity Markets'. In: *Proceedings of the VII Symposium of Specialists in Electric Operational and Expansion Planning*, Curitiba (Brasil), May 21 - 26.

Oren, S.S., Spiller, P.T., Varaiya, P.V. and Wu, F.F. 1995. 'Nodal Prices and Transmission Rights: a Critical Appraisal'. *The Electricity Journal* 8 (3): 24-35.

Overbye, M. 2000. *Norway's government resigns after losing power plant vote*. CNN, March 9. Obtained from: http://europe.cnn.com/2000/WORLD/europe/03/09/norway.govt.02.

Pacific Gas and Electric (2000) *Annual Report*. Obtained from: www.pgecorp.com/financial/reports/pdf/FS_(2000)final.pdf.

Pearce, D.W. and Turner, R.K. 1990. *Economics of Natural Resources and the Environment*. Baltimore, The Johns Hopkins University Press.

Pérez-Arriaga, I.J. 2003. personal communication, August 14.

Pérez-Arriaga, I.J. 2001. *Long-term reliability of generation in competitive wholesale markets, a critical review of the issues and alternative options*. IIT Working Paper IIT-00-098IT, Madrid, Universidad Pontificia Comillas, Instituto de Investigación Tecnológica. Obtained from: www.iit.upco.es/docs/01JIPA2001.pdf.

Pérez-Arriaga, I.J. and Meseguer, C. 1997. 'Wholesale marginal prices in competitive generation markets'. *IEEE Transactions of Power Systems* 12 (2): 710-717.

Pfeifenberger, J.P. and Tye, W.B. 1995. 'Handle with care: A primer on incentive regulation'. *Energy Policy* 23 (9): 769-779.

PG&E (Pacific Gas and Electric) 2000. *Annual Report*. Obtained from: www.pgecorp.com/financial/reports/pdf/FS_2000final.pdf.

PJM Interconnection, L.L.C. 2003. *Reliability Assurance Agreement Among Load Serving Entities in the MAAC Control Zone, Second Revised Rate Schedule FERC No. 27*. Obtained from www.pjm.com.

Poza, E. and Ackermann, T. 2001. 'Centralised Power Generation versus Distributed Power Generation: a System Analysis'. In: *Proceedings, First International Symposium on Distributed Generation: Power System and Market Aspects,* Stockholm, June 11-13.

Regulation (EC) No 1228/2003 of the European Parliament and of the Council of 26 June 2003 on

conditions for access to the network for cross-border exchanges in electricity, OJ 2003 L 176: 1-10.

Regenesys Technologies 2003. web site: www.regenesys.com.

Roberts, L. and Formby, R. 2001. 'Market Participant Experiences with Demand Side Bidding and Future Direction'. In: *Proceedings of the Market Design 2001 Conference*, Stockholm 7 - 8 June 2001: 67-76.

Roe, E., Van Eeten, M., Schulman, P. and De Bruijne, M. 2002. *California's Electricity Restructuring, The Challenge to Providing Service and Grid Reliability*. Concord (CA), EPRI.

Rosenberg, A.E. 2002. 'Congestion Pricing or Monopoly Pricing?'. *The Electricity Journal* 13 (3): 33-41.

Ross, S.A., Westerfield, R.W. and Jaffe, J.F. 2002. *Corporate Finance*. New York, McGraw-Hill (6^{th} ed.).

Sep and EnergieNed (the Dutch Electricity Generating Board and the Associations of Energy Distribution Companies in the Netherlands) 1999. *Electricity in the Netherlands 1998*. Arnhem, EnergieNed.

Sep (the Dutch Electricity Generating Board) 1987. *Electricity in the Netherlands 1986*. Arnhem, Sep.

Sæle, H. and Grønli, H. 2001. 'Small customers as active peak power providers in periods of capacity problems'. In: *Proceedings of the Market Design 2001 Conference,* Stockholm 7 - 8 June: 77-84.

Schweppe, F.C. 1978. 'Power Systems '2000': Hierarchical Control Strategies'. *IEEE Spectrum*, July: 42-47.

Sheffrin, A. 2002. 'California Power Crisis: Failure of Market Design or Regulation?'. *IEEE Power Engineering Review* 22 (8): 8-11.

Shuttleworth, G. 1997. 'Getting Markets to Clear'. Letter to the Editor, *The Electricity Journal* 10 (3): 2.

Shuttleworth, G., Falk, J., Meehan, E., Rosenzweig, M. and Fraser, H. 2002. *Electricity Markets and Capacity Obligations, A Report for the Department of Trade and Industry*. London, NERA.

Skantze, P. and Ilic, M.D. 2001. 'Investment Dynamics and Long Term Price Trends in Competitive Electricity Markets'. In: *Proceedings, IFAC Symposium on Modeling and Control of Economic Systems,* Klagenfurt, Austria, September 6-8.

Spence, A.M. 1977. 'Entry, capacity, investment and oligopolistic pricing'. *Bell Journal of Economics* 8: 534-544.

Stoft, S.E. 1999. *How to Provide Locational Signals for Generator Investment in the Absence of Congestion Pricing*. Mimeo. Obtained from: http://stoft.com/metaPage/lib/Stoft-1999-Non-LMP-signals-for-Gen.pdf.

Stoft, S.E. 2000. *PJM's Capacity Market in a Price-Spike World*. Working paper PWP-077, Berkeley, University of California Energy Institute.

Stoft, S.E. 2002. *Power System Economics: Designing Markets for Electricity*. Piscataway (NJ), IEEE Press.

Tabors, R.D. 1999. *Transmission Pricing in PJM: Allowing the Economics of the Market to Work*. TCA Working Paper 0299-0216.

TenneT 2004. Web site: www.tennet.nl; data on installed capacity: www.tennet.nl/overige/030_productiegegevens.

Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, Massachussetts Institute of Technology.

Tugwell, F. 1988. *The Energy Crisis and the American Political Economy, Politics and Markets in the Management of Natural Resources*. Stanford, Stanford University Press.

Turvey, R. 2000. 'Infrastructure access pricing and lumpy investments'. *Utilities Policy* 9: 207-218.

UCTE (Union for the Co-ordination of Transmission of Electricity) 2002a. *UCTE Power Balance Forecast 2002-2004*. Brussels, UCTE. Obtained from: www.ucte.org/publications/library/e_default_2002.asp.

UCTE (Union for the Co-ordination of Transmission of Electricity) 2002b. *UCTE System Adequacy Forecast 2003 – 2005*. Brussels, UCTE.

UCTE (Union for the Co-ordination of Transmission of Electricity) 2002c. *UCTE Power Balance Retrospect 2001*. Brussels, UCTE. Obtained from: www.ucte.org/pdf/Publications/2001/Retrospect_2001.pdf.

UCTE (Union for the Co-ordination of Transmission of Electricity) 2003. *UCTE System Adequacy Forecast 2004 – 2010*. Brussels, UCTE. Obtained from: www.ucte.org/publications/library/e_default_2004.asp.

Union of Concerned Scientists 2000. *Public Utility Policy Act briefing*. Obtained from: www.ucsusa.org/energy/brief.purpa.html.

United States Congress 1978. *Public Utility Regulatory Policy Act (PURPA)*. Obtained from: www.ferc.gov.

United States Congress 1992. *Energy Policy Act (EPact)*. Obtained from: http://energy.nfesc.navy.mil/docs/law_us/92epact/hr776toc.htm.

University of California Energy Institute, California Electricity Market Data web site: www.ucei.berkeley.edu/ucei/datamine/datamine.htm.

Utilities (Dutch monthly publication) 2003. Vol. 4, p. 47 (list of planned generation projects).

Van Eck, T., Rödel, J.G. and Verkooijen, A.H.M. 2002. 'Binnenlands vermogen biedt onvoldoende zekerheid'. *Energietechniek* 9: 40-43.

Van Werven, M. 2003. *Monitoring van voorzieningszekerheid: hoever reikt het vermogen van de elektriciteitsmarkt?* Master's thesis, Delft University of Technology.

Vázquez, C., Rivier, M. and Pérez-Arriaga, I.J. 2000. *On the use of pay-as-bid auctions in California, some criticisms and an alternative proposal*. IIT working paper IIT-00-077A, Madrid, Universidad Pontificia Comillas, Instituto de Investigación Tecnológica.

Vázquez, C., Rivier, M. and Pérez-Arriaga, I.J. 2002. 'A market approach to long-term security of supply'. *IEEE Transactions on Power Systems* 17 (2): 349-357.

Vázquez, C., Batlle, C., Rivier, M. and Pérez-Arriaga, I.J. 2004. 'Security of supply in the Dutch electricity market: the role of reliability options'. Paper presented at the *Conference on "Competition and Coordination in the Electricity Industry",* Toulouse, January 16 - 17.

Visudhiphan, P., Skantze, P. and Ilic, M. 2001. 'Dynamic Investment in Electricity Markets and Its Impact on System Reliability'. In: *Proceedings of the Market Design 2001 Conference,* Stockholm, 7and 8 June: 91-110.

Von der Fehr, N.-H. M. and Harbord, D.C. 1997. *Capacity Investment and Competition in Decentralised Electricity Markets*. Mimeo, University of Oslo, Department of Economics.

Von der Fehr, N.H. M and Harbord, D.C. 1998. *Competition in Electricity Spot Markets, Economic Theory and International Experience*. Obtained from: http://faculty-gsb.stanford.edu/wilson/E542/documents/Electricity/Harbord,Survey.pdf.

Watts, P.C. (pseudonym) 2001. 'Heresy? The Case Against Deregulation of Electricity Generation'. *The Electricity Journal* 14 (4): 19-24.

Weare, C. 2003. *The California Electricity Crisis: Causes and Policy Options*. San Francisco, Public Policy Institute of California. Obtained from: www.ppic.org/content/pubs/R_103CWR.pdf.

Willis, K.G. and Garrod, G.D. 1997. 'Electric Supply Reliability, Estimating the Value of Lost Load'. *Energy Policy* 25 (1): 97-103.

Wolak, F.A. and Patrick, R.H. 1997. *The Impact of Market Rules and Market Structure on the Price Determination Process in the England and Wales Electricity Market*. Mimeo. Obtained from: ftp://zia.stanford.edu/pub/papers/eandw.pdf.

Wong, W., Chao, H., Julian, D., Lindberg P. and Kolluri, S. 1999. 'Transmission Planning in a Deregulated Environment'. In: *1999 IEEE Transmission and Distribution Conference Proceedings,* Vol. 1: 350-355.

Woolfolk, J. 2001. 'Deregulation overlooked long-term power buying'. *San Jose Mercury News*, Jan. 11. Obtained from: www0.mercurycenter.com/.

World Bank 2001. *The California Power Crisis: Lessons for Developing Countries*. Washington, DC, The World Bank, Energy and Mining Sector Board.

Wu, F.F., Varaiya, P., Spiller, P. and Oren, S.S. 1996, 'Folk Theorems on Transmission Access: Proofs and Counterexamples'. *Journal of Regulatory Economics* 10: 5-23.

Yardley, J. 2001. 'Texas Learns in California How Not to Deregulate'. *The New York Times,* January 10.

Appendix: A dynamic model of several capacity mechanisms

A.1 Introduction

To demonstrate the dynamic effects of several capacity mechanisms, a simple model was constructed in Microsoft Excel. The purpose of the model is to gain an understanding of how imperfect investment behavior (for instance due to risk averse investment behavior, or due to imperfect information) can be corrected through capacity mechanisms. Therefore a model of an energy-only market is compared to a system with capacity payments, operating reserves pricing and a system with a capacity requirement. Reliability contracts were not modeled but under the assumptions of the model they can be expected to behave similarly to a capacity requirement. The main difference between reliability contracts and capacity requirements is their robustness against the abuse of market power. Because the model assumes perfectly competitive behavior, this difference will not be apparent.

The goal of the model is to gain understanding of the factors that influence the dynamics of the capacity mechanisms, in particular the extent to which the different capacity mechanisms are robust against investment decisions that are not socially optimal. Predicting market developments is not a goal, as opposed to Ford (1999 and 2001). Structurally, the model in this Appendix resembles Ford's dynamic models. In considering different types of investment behavior, use was also made of Visudhiphan (2001) and Botterud et al. (2002). Hobbs et al. (2001c) present an equilibrium model of the same capacity mechanisms as this appendix. They conclude that operating reserves pricing and capacity requirements can result in the same level of reliability, at the same cost, and with the same mix of different types of generator technology, and that both options may improve system adequacy.

As the model is simple, it does not provide an accurate description of how the modeled systems would work in practice but only a first approximation. Market power, for instance, is not modeled: neither short-term strategic behavior such as capacity withholding, nor long-term strategic considerations such as entry deterrence. Nevertheless, the model provides interesting insights in the way different capacity mechanisms may be able to dampen investment cycles.

The model shows the dynamic development of a fictive electricity system (loosely based

upon the Dutch electricity system) for four different capacity mechanisms: an energy-only market, capacity payments, operating reserves pricing and a system with capacity requirements. The model calculates the equilibrium volume of generation capacity for several periods per year for a number of successive years. Demand is assumed to grow uniformly. For each year, prices are calculated, from which an investment incentive is derived. After a delay, the new generation capacity becomes available for production. Thus higher prices lead to new generation capacity with a delay.

The investment signal is calibrated so the runs of the model all produce long-run average generator revenues equal to the long-run marginal cost. (The reasons will be explained below.) As the model will therefore not show chronic over or under investment, different criteria need to be applied to evaluate the merits of the different capacity mechanisms. Following the analysis in Chapters 5 and 6, a capacity mechanism will be considered successful if it stabilizes investment, i.e. it keeps power shortages to a minimum and stabilizes electricity prices.

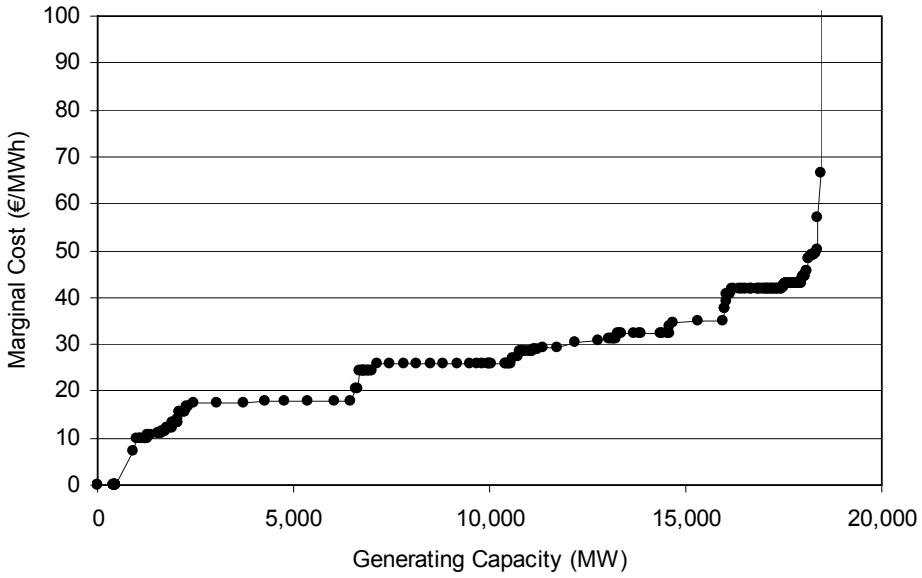


Figure A.1: Supply Function

A.2 Assumptions

Supply curve

The supply curve is based upon data on the marginal cost of generation of Dutch

generators (Figure A.1).[70] Outside the figure – for demand realizations in excess of available generation capacity – the supply function jumps to 2500 €/MWh for the volume of interruptible contracts (which is arbitrarily set to 500 MW), after which it jumps again to 8600 €/MWh, the estimated average value of lost load in the Netherlands (Bijvoet et al., 2003). Strictly speaking, interruptible contracts are part of the demand function. However, the effect of a controllable reduction of demand is similar to an equivalent increase in generator output. It greatly simplifies the model to assume a perfectly price-inelastic demand function and to model the interruptible contracts as additional generation resources.

Table A.3: Load-Duration Data

	Hours	Hours (cumulative)	Demand (MW)
Winter super peak	50	50	16,201
Winter peak	704	754	15,057
Spring/Autumn super peak	100	854	15,043
Summer super peak	50	904	14,222
Spring/Autumn peak	1429	2333	13,744
Summer peak	720	3053	13,504
Winter should	704	3757	11,719
Spring/Autumn shoulder	1429	5186	11,196
Summer shoulder	720	5906	10,777
Winter off peak	704	6610	9,426
Summer off peak	720	7330	9,028
Spring/Autumn off peak	1429	8759	8,952

A proper load-duration curve could not be obtained. (It may not be available at all for the Netherlands, for the reasons discussed in Section 5.1.3.) Instead, aggregated load-duration data were used which are shown in Table A.3.[71] For three seasons (winter, summer, and spring/autumn), average load is given for four periods each (off peak, shoulder, peak and super peak). Twelve load segments result, which are grouped not by time but by (average) load. A crude load-duration curve results, which is plotted in Figure A.2.

Perfect competition

The model assumes perfect competition: prices equal the marginal cost of generation, unless available generation capacity is less than demand. In that case, prices jump first to the price of interruptible contracts (see below). When these are exhausted, price jump further to the average value of lost load.

[70] Courtesy of ECN (Energy Research Centre of the Netherlands). Special thanks to Maroeska Boots, who provided an algorithm to calculate the intersections between the demand realizations and the supply function.
[71] Also courtesy of ECN.

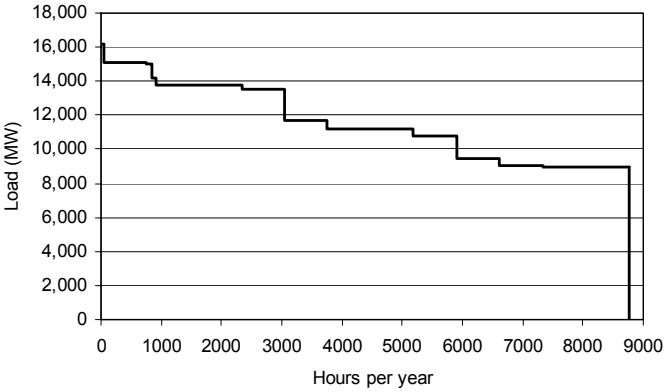


Figure A.2: Load-Duration Data

Demand growth

Initially, demand is assumed to increase with a constant growth rate, so the load-duration curve of a given year can simply be found by multiplying the previous year's load-duration curve by the demand growth rate. In Section A.4.6, the effect of a change in the growth rate of demand will be modeled.

Demand price-elasticity

Demand is assumed to be perfectly price-inelastic for most cases. An exception is that a volume of interruptible contracts is assumed for a price P_c of € 2500/MWh, so at that level there is some demand elasticity. This is modeled as additional generation capacity with a marginal cost of € 2500/MWh. This price was chosen with the experience of the summer of 2003 in mind, when it was deemed necessary to raise the price cap at the APX spot market to this level because that would allow additional resources to become available. The volume of interruptible contracts V_{ic} is set at 500 MW.

Generator availability

Generator availability is assumed to be 100%. In reality, maintenance and forced outages, which occur unpredictably, lower the availability. This increases the uncertainty regarding the required volume of installed generation capacity. As more uncertainty is detrimental to investment, this will probably only exacerbate the tendency towards investment cycles. Modeling availability as 100% therefore is a safe assumption in this case.

Generator construction lead time

It is assumed that it takes five years between the decision to build new generation capacity and the time it becomes available.[72] This period cannot be varied in the model.

[72] This reflects the long permitting time in many countries, such as the Netherlands.

However, the dynamic effects of a shorter lead time are likely the same as those of a lower growth rate, except that the time scale changes. A shorter lead time means that high prices lead more quickly to new generation capacity, so the probability of the development of a shortage is reduced. At a lower average growth rate of demand, there is more time between the first occurrence of high prices (when the interruptible contracts set the price) and the development of a shortage, so the probability of the development of a shortage is also reduced.

Imperfect information

It is assumed that generating companies are not able to forecast the demand for generation capacity accurately five years into the future, when new capacity becomes available. This assumption is modeled by letting the generating companies underestimate the growth rate of demand by a certain percentage. On the other hand, it is assumed that a part of new investment is inspired not by forecasts of supply and demand but by current prices. Therefore price spikes lead to investment (which may exceed the need for new capacity). The way the investment signal is modeled will be explained in detail in Section A.3.3.[73] The assumptions regarding investor's response to electricity prices are similar to those that Ford (1999) uses. An important difference is that in our model it is assumed that generating companies know how much new generation capacity their competitors currently are constructing. This will be called capacity 'in the pipeline'. This information should be available because, for instance, permitting processes are public. Investors consider the volume of generation capacity in the pipeline in their investment decisions. As a result, they may decide not to construct new generation capacity even if there is a strong price signal, if there already is much new capacity under construction.

Type of new generation capacity

For the sake of simplicity, all new capacity is assumed to be modern combined cycle gas technology, with fixed costs C_f of 86,740 €/MW per year and variable (fuel) costs C_v of 23.81 €/MWh. These figures are based upon an estimate of the costs of the Shell/Intergen project, the largest new generator under construction in the Netherlands. (Source: Utilities, April 2003). This means that in the model, new capacity will be medium load capacity, so the increased peak demand will be served by existing generators, who are moved upwards on the supply curve. A consequence is that the middle section of the supply curve is 'flattened' by the addition of identical generators with identical variable costs. This could increase generator revenue volatility, as the price would be equal to the variable cost of an increasing number of generators for an increasing amount of time. It may be that investors indeed would choose this strategy (see Section 5.4.5) but it would be limited by the need for daily peaking capacity and capacity with a high ramp speed. Therefore a certain mix would always be expected. However, as decommissioning is not

[73] Future research could simulate cases where demand growth is uncertain and represented probabilistically. An example was given in Figure 5.10 on page 95. If investment behavior is based upon an extrapolation of recent experience, e.g. the average growth rate of the last five years, periods of low demand growth will result in an under-estimation of the future demand and hence in too little investment. (This would be comparable to the 'backward looking investment' in the model of Visudhiphan et al. (2001).)

modeled (see below), the model does maintain a certain mix of plants with different variable costs.

Free entry

It is assumed that any party can construct a new generation plant as easily and at the same cost as existing generating firms. This means that a certain amount of investment may take place also when there is no shortage if the long-run marginal cost of generation is below the average electricity price.

Replacement investments

Aging and retirement of plants are not modeled. This implies an assumption that retired plants are always replaced in time. This results in an over-estimation of episodes of over-investment. In reality, plants that would retire during these episodes would not be replaced (and they might also be retired earlier), which would lead to less excess capacity. Because the purpose of the model is to assess possibilities to avoid construction cycles, the fact that the aftermath of a construction boom is not modeled accurately is not so important. It should be remembered, however, that this will influence the price: an over-estimation of available capacity after a boom will lead to an under-estimation of the electricity price during those years, as the marginal generation cost determines the price.

A.3 Model structure

A.3.1 Electricity price calculation

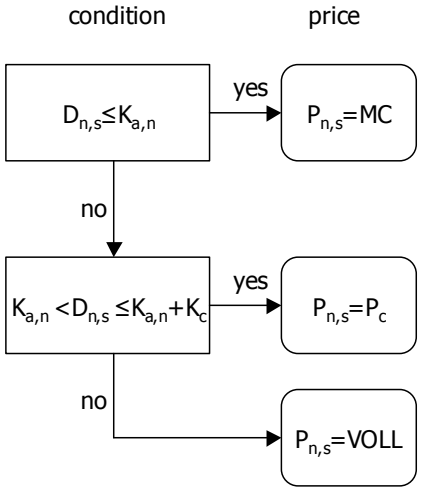
Demand

For each of the 12 periods in Table A.3, demand D_{n+1} in year n+1 is found by multiplying the demand in the previous year D_n with 1 plus the annual growth rate g:

$$D_{n+1}=D_n(1+g) \qquad (A.1)$$

Energy-only and capacity requirement

The model calculates electricity prices based upon supply and demand equilibria for each of the 12 periods in Table A.3, for each of the years 2004 through 2030. Demand is assumed to be price-inelastic, and is therefore represented by a single figure. The basic supply function is given in Figure A.1 but is expanded with new generation capacity. The market price normally is equal to the marginal cost of generation. If demand exceeds available generation capacity, the price is equal to the cost of interruptible load contracts P_c or, if this resource is also exhausted, equal to the average value of lost load. In theory, this should be the level at which customers stop purchasing electricity. (See Section 5.2.4.) It assumes that there is no administrative price cap in place below the average value of lost load.



$D_{n,s}$ = demand in section *s* of the load-duration curve in year *n*.

$K_{a,n}$ = available generation capacity in year *n*.

Kc = volume of interruptible contracts

$P_{n,s}$ = electricity price when load is in section *s* of the load-duration curve in year *n*.

P_c = interruptible contract price

MC marginal cost of generation

VOLL the average value of lost load

Figure A.3: Price determination algorithm for an energy-only market

For each year *n*, the equilibrium price $P_{n,s}$ is calculated for the 12 different periods listed in Table A.3. In principle, the price is equal to the marginal cost of generation. Because there is no real supply function but only a collection of data points, the model interpolates between these points. If demand exceeds supply, the price is set by the price of interruptible contracts. These are modeled as additional supply, only at a much higher price. If the interruptible contracts are exhausted, the price is set equal to the average value of lost load. Figure A.3 schematically represents the algorithm used to determine the price.

Operating reserves

In an operating reserves market, price calculation is slightly more complex. In most of the runs presented here, the price of interruptible capacity is higher than the price of the operating reserves. This possibility is represented by the right-hand side of Figure A.4. For periods in which demand $D_{n,s}$ is smaller than available generation capacity $K_{a,n}$ minus the volume of capacity that is contracted as operating reserves K_{or}, the price $P_{n,s}$ is equal to the marginal cost *MC*:

$$If\ D_{n,s} \leq K_{a,n} - K_{or},\ then\ P_{n,s} = MC \qquad (A.2)$$

If demand can only be met by using the operating reserves, the operating reserves price P_{or} determines the market price:

$$If\ \ K_{a,n} - K_{or} < D_{n,s} \leq K_{a,n},\ then\ \ P_{n,s} = P_{or} \qquad (A.3)$$

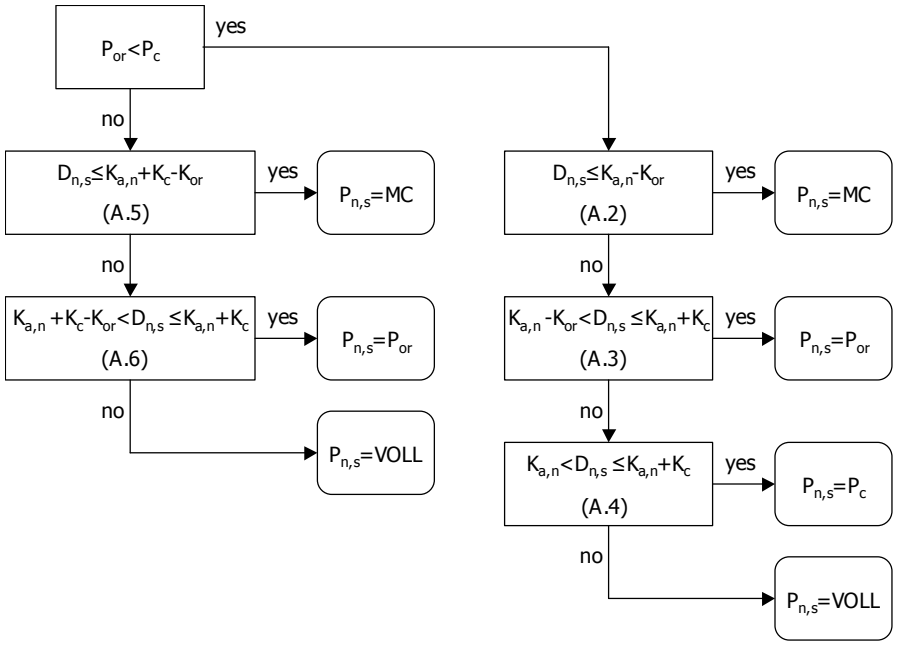


Figure A.4: Price determination algorithm for a market with operating reserves pricing

The model makes a simplifying approximation, as actually the price at which generating companies are indifferent whether they sell operating reserves to the system operator or energy to the market is given by $P_{or}+MC$. The relative error is small, however, as P_{or}=2000 €/MWh in most runs, whereas the highest variable cost in the system is 67 €/MWh.

P_{or} is determined by the system operator, who estimates its optimal value in combination with the volume of operating reserves K_{or} that he contracts.[74]

When the operating reserves are exhausted, the interruptible contracts will be called upon, so they set the price:

$$If \;\; K_{a,n} < D_{n,s} \leq K_{a,n} + K_c, \; then \;\; P_{n,s} = P_c \qquad (A.4)$$

If demand exceeds the volume of available generation capacity plus the volume of interruptible contracts, there will be a shortage and the price rises to the average value of lost load.

The model also allows for the possibility that the price of interruptible contracts is below the price of the operating reserves, which is represented by the left-hand side of Figure

[74] See 6.4 for a brief explanation of how the operating reserves price and volume are determined, or Stoft (2002) for a more detailed description.

A.4. In this case, the interruptible contracts are included in the operating reserves as extra generation capacity.

$$If\ D_{n,s} \le K_{a,n} + K_c - K_{or},\ then\ P_{n,s} = MC \qquad (A.5)$$

If demand can only be met by using the operating reserves, the operating reserves price determines the market price:

$$If\ \ K_{a,n} + K_c - K_{or} < D_{n,s} \le K_{a,n} + K_c,\ then\ \ P_{n,s} = P_{or} \qquad (A.6)$$

Again, the same simplification was made as in (A.3): actually, $P_{n,s}$ would be equal to P_{or}+MC, but the latter term is relatively small.

As the operating reserves include the interruptible contracts, there is no other resource when they are exhausted, so scarcity develops and the price becomes equal to the average value of lost load.

A.3.2 Investment in energy-only and operating reserves markets

The most difficult part of the model is the question how to model investment decisions. Starting point is the assumption that generating companies have incomplete information about the development of demand. Therefore, it is assumed that they base their investment decisions partly upon a projection of future demand and partly upon price information. The model does not distinguish between different types of investors. Therefore it uses only one investment signal S_n (measured in MW). This signal is composed of a price-related signal $S_{p,n}$, a signal based upon a forecast of peak demand $S_{f,n}$ and a component $S_{c,n}$, which reflects any payments to generating companies as a result of a capacity mechanism (capacity payments, payments for operating reserves or payments for capacity credits, depending on the case).[75] Note that the signal is not equal to these payments but a function of them. In an energy-only market, the $S_{c,n}$ equals zero (or is small, if payments are made for a small operating reserve for the purpose of operational control of the system).

The investment signal can be capped to avoid excessive reactions to shortages. Even if prices are extremely high, it is not to be expected that investors would choose a higher level of investment than, say, 50% of existing capacity. In most runs of the model, this arbitrarily chosen 'common sense' limit is not reached. Thus the investment signal S_n is given by:

$$S_n = MIN\left\{(S_{p,n} + S_{f,n} + S_{c,n}), (50\% \cdot K_{a,n})\right\} \qquad (A.7)$$

The investment *signal* is not necessarily the same as the actual investment because investors also consider the volume of generation capacity already in the pipeline, as will

[75] Ford (2001) also uses an investment signal based upon a mix of price information and information related to the capacity margin.

be described below.

The demand forecast-based component of the investment signal

Investors estimate the future demand for new generation capacity by comparing existing capacity to their projection of peak electricity demand in five years. The purpose of this model is to assess how different capacity mechanisms are able to stabilize investment if investors make sub-optimal decisions. Therefore the model has the option to add an error to the demand forecasts, which causes investment to lag demand growth so a shortage develops over time. When a shortage develops, high prices provide a strong investment signal.

Because the model works with an investment delay of five years, it is assumed that investors try to forecast demand five years into the future when they consider the need for new plant. Their forecast of peak electricity demand $D_{f,n+5}$ for year $n+5$ is:

$$D_{f,n+5} = D_n \cdot (1 + g + e)^5 \qquad (A.8)$$

The variable e is the demand estimation error, which can be included to simulate underestimation of demand or risk aversion. It has a negative value. In effect, it means that investors project demand with a growth rate of $(g+e)$, rather than the real growth rate of g. (In the base case, g=2.5% and e=-1.0%, so the investors forecast demand to grow at a rate of 1.5% per year.)

The error is introduced to compensate for the fact that demand growth is constant in the model. Without an error, the investment decisions are perfect in this model and the volume of available generation capacity always is just enough to meet peak demand. In the real world of variations in the growth rate of demand, this would result in periods of scarcity, followed by investment booms and periods of excess capacity. In the model, the error leads to periods of scarcity. Without the error, the effectiveness of the capacity mechanisms could not be analyzed. The error can also be interpreted as representing risk-aversion: by under-estimating demand growth, investors reduce the risk that they invest too much and create excess capacity, which would lead to prices below total cost in a competitive market. As investment is driven by two components, prices and the demand forecast, the demand estimation error is a means of making investment more or less dependent upon prices relative to the demand forecast.

The investment signal that is derived from the demand forecast S_f (MW) is defined as the difference between the demand forecast $D_{f,n+5}$ and the existing volume of generation capacity $K_{a,n}$, if this difference is positive, and zero otherwise:

$$S_{f,n} = MAX(D_{f,n+5} - K_{a,n}, 0) \qquad (A.9)$$

The price-related component of the investment signal

The price signal component of the investment signal is based upon the operating profit that a new generating plant would have made, given existing electricity prices. The larger

the potential operating profit for a new plant, the stronger the investment signal. Even without an immediate need for new capacity, the model creates a low investment signal because new units are cheaper than some of the existing ones. The high prices during episodes of scarcity provide a much stronger investment signal. Thus, if a period of scarcity was not anticipated (due to the error in forecasting demand), the price-based investment signal creates a correction (though it may be too late to avoid a period of scarcity).

To arrive at the price signal in year n, the potential operating profit of new a new plant $\Pi_{n,s}$ per unit of generation capacity is calculated for each load segment s in the load-duration data of Table A.3. $\Pi_{n,s}$ is equal to the difference between the price P and the marginal cost C_v of new generation capacity, multiplied by the number of hours h_s that this price occurs in a year, if this is positive; otherwise it is zero.

$$\Pi_{s,n} = MAX\{(P_s - C_v) \cdot h_s, 0\} \qquad (A.10)$$

The number of hours h_s that a price P_s occurs can be found with the load-duration data. h_s is the duration (in hours) of that segment of the load-duration curve. The units of Π are €/MW per year.

The total potential operating profit of new plant in a given year Π_o is the sum of the operating profits for each section of the load-duration curve $\Pi_{o,s}$:

$$\Pi_n = \sum_s \Pi_{o,s} \qquad (A.11)$$

The price component of investment signal $S_{p,n}$ (MW) is derived from the potential operating profit by multiplying it with a scaling factor F_i (MW^2/€):

$$S_{p,n} = \Pi_n \cdot F_i \qquad (A.12)$$

This factor is chosen so the average prices are close to the long-run marginal cost of generation, which is estimated at about 42 €/MWh, as much as possible. (It will be indicated when this is not the case: under certain extreme assumptions, the price may be consistently lower, for instance.) The factor is adjusted each run, reflecting the assumption that structurally different market conditions lead to a different investment reaction to price signals. For instance, if the demand estimation error e is set low, there generally will be sufficient capacity, which depresses prices. If prices are below the long-run marginal cost for many years in a row, it is unlikely that there will be much new investment, so in that case F_i should be set lower.

Ideally, F_i would be an endogenous factor, which would reflect the actual response of investors to price signals. In the absence of empirical data and for lack of a better alternative, F_i is calibrated so generating companies' long-run average revenues equal their costs. This procedure assumes that investors know the long-run market conditions, which is in contradiction with the assumption that they have imperfect foresight. However, the sensitivity analysis will show that the basic results of the model –

investment cycles in an energy-only market – occur for a wide range of values of F_i. The reason is that a lower investment factor will cause too little reaction when prices start to rise, so a shortage develops, while a higher investment factor will cause an over-reaction, if a shortage develops. In other words, the runs of the model will show that given imperfect foresight and the lead time for new generation capacity, there is no perfect investment response to prices that eliminates the tendency towards investment cycles. As the purpose of the model is to test the ability of different capacity mechanisms to dampen investment cycles, the analysis is therefore not highly sensitive to the choice of F_i.

The investment signal from operating reserves payments

By contracting operating reserves, the system operator provides an additional revenue stream to the generating companies. The payments $S_{c,n}$ are determined by an algorithm similar to the price determination algorithm that is depicted in Figure A.4. If available generation capacity is large enough that the system operator can contract his desired volume of reserves, the electricity price is determined by the marginal cost of generation MC and the price of the reserves is determined by the marginal costs of keeping generation capacity available, which is estimated to be 10 €/MWh. This is about equal the assumed annual fixed costs of new generation capacity C_f, divided by the number of hours in a year. At times when the market price is determined by the willingness to pay of the system operator, P_{or}, the payments for the operating reserves capacity also equal P_{or}. When there is insufficient generation capacity to meet demand, the reserves are exhausted and the system operator makes no payments. (In reality, the system operator will need to continue to purchase a small volume of reserves for the purpose of maintaining operational stability, like he does in an energy-only market as well. In the model, this effect will be disregarded.)

The investment signal from capacity payments

Only static capacity payments are modeled. They simply are a reimbursement for available generation capacity. The capacity payments are equal to the total volume of installed capacity multiplied by the payment per unit of capacity.

The investment decision

The model assumes that while investors have imperfect information about the future development of supply and demand, they do know the volume of generation capacity that is under construction. They incorporate this knowledge in their investment decision by subtracting the volume in the pipeline from the investment signal. The volume of capacity under construction $K_{c,n}$ ('in the pipeline') in year n is the sum of the investment decisions of the previous four years. Earlier investment decisions have already been realized and are included in the volume of available capacity.

$$K_{c,n} = K_{i,n-4} + K_{i,n-3} + K_{i,n-2} + K_{i,n-1} \qquad (A.13)$$

Thus the volume of generation capacity $K_{i,n}$ that actually is invested in, in year n, is equal to the investment signal S_n minus the volume of generation capacity that is in the pipeline

$K_{c,n}$ in that year, if this difference is greater than zero. If the capacity under construction is greater than the investment signal, there is no investment.

$$K_{i,n} = MAX(S_n - K_{c,n}, 0) \qquad (A.14)$$

S_n is given by equation *(A.7)*.

A.3.3 Investment in a system with capacity requirements

In a market with capacity requirements, the investment signal needs to be modeled differently than in an energy-only market or a market with operating reserves pricing. The explicit purpose of capacity requirements is to provide a clearer and more stable investment signal than the energy price, so it would be a misrepresentation of the system to model investment as still being driven by the energy prices alone. Capacity requirements improve market transparency by creating a predictable demand for generation capacity (for the planning horizon of the regulator who determines the levels of the capacity requirements).

While the investment signal necessarily is different, some basic features are the same as in the models of the energy-only market and the operating reserves pricing market. The delay with which new capacity arrives in the market is five years and investors have imperfect foresight. If a shortage develops, investors react to high energy prices the same way as in an energy-only or operating reserves market. However, below it will be shown that because the capacity requirement reduces the frequency of shortages and near-shortages significantly, the effect of this price signal will be smaller in the model.

The assumption will be made that, in addition to the price signal that exists in an energy-only market, generation investment is driven by the expected demand for capacity credits. The load-serving entities will have a demand for capacity equal to the capacity requirement, enforced by the penalty that is higher than the long-run marginal cost of generation. A second assumption is that aggregate load data are published and that the regulator publishes his expectation of future capacity requirements or, even better, that the capacity requirements are established several years in advance. As a result (if the regulator does his work well), the future development of the demand for capacity is more predictable.

Thus the investment signal S_n in year n consists of the price signal $S_{p,n}$ and the forecast $S_{f,n}$ (which in this case is based upon the regulatory requirement for generation capacity, rather than forecast peak demand). These incentives are structured as follows.

Price signal

The price signal $S_{p,n}$ is the same as in an energy-only market, which is given by equations *(A.10)* through *(A.12)*. In the runs of this model, this signal turns out to play a minor role, as most generator revenues are from the sales of capacity credits.

Forecast of demand for capacity

The generating companies project the future demand for generation capacity by making a demand projection. They forecast the total required volume of generation capacity and compare this to the existing volume of generation capacity to arrive at an estimate of the demand for new capacity. The generating companies extrapolate current demand D_n, assuming a fixed growth rate. As in the energy-only market model, the growth rate is g% per year and there is an option to include an error of e percentage points to this rate (the same as in (A.8)). The generating companies' demand forecast $D_{f,n+t}$ for in year $n+t$ is given by:

$$D_{f,n+t} = D_n (1 + g + e)^t \qquad (A.15)$$

The forecast of the required volume of generation capacity $K_{r,n+t}$ in year $n+t$ is equal to the demand forecast, made in *(A.15)*, multiplied by 1 plus the regulator's reserve requirement r (measured in percent of peak demand).

$$K_{r,n+t} = (1 + r) D_{f,n+t} \qquad (A.16)$$

The projected volume of generation capacity is compared to the existing volume of generation capacity $K_{a,n}$. The difference is the expected need for new investment:

$$S_{f,n} = K_{r,n+t} - K_{a,n} = (1 + r) D_{f,n+t} - K_{a,n} \qquad (A.17)$$

Forecasts of the demand for generation capacity should be made for at least five years in advance because in the investment decision the forecasts are compared to the volume in the pipeline, which may take that long to realize. In the runs shown here, all forecasts are for 5 years in advance (so t=5).

The investment decision

The total investment signal S_n is found by adding (A.12) and (A.17):

$$S_n = MIN\{(S_{p,n} + S_{f,n}), (50\% \cdot K_{a,n})\} \qquad (A.18)$$

The actual amount that is invested again depends upon the volume of generation capacity in the pipeline is the same as in the case of an energy-only market; see equation *(A.14)*.

A.3.4 Presentation of model output

For each year in the period 2004-2030 the model calculates the development of demand, generator revenues (for the 12 periods per year) and generation capacity. These results are summarized in graphs with the years on the X-axis, electricity demand and generation capacity on the first (left) Y-axis and generator revenues on the second (right) Y-axis.

Generator revenues per unit of electricity output are defined as follows. In the case of an

energy-only market, they are equal to the annual average electricity price for the entire market (total payments for electricity divided by the total volume of electricity sold). In the case of an operating reserves pricing market, the average generator revenues are calculated as the sum of the payments for electricity plus the operating reserves payments, divided by the total volume of energy sold. In a market with a capacity requirement, generator revenues are equal to total electricity payments plus the capacity payments, divided by the total volume of energy sold.

Next to each graph the values of the main parameters are shown. While in principle only one parameter is changed per run, the value of the investment factor F_i is adjusted in each run in order to mimic the adjustment of investment behavior to structurally different market conditions. Unless indicated otherwise, the long-run average generator revenues therefore are 42 €/MWh. If not, the value of F_i is stated next to the graph.

In addition it is shown in which years the volume of generation capacity plus interruptible contracts is insufficient to meet all demand and therefore service interruptions need to be imposed. The latter is an approximation, due to the inaccurate load-duration curve data. As the load-duration curve data consist of average figures for 12 periods of load, short episodes of scarcity may not show up. A stable market with an optimal duration of load shedding of only a few hours per year, for instance, would be represented as a system without any outages in this model. This would wrongly suggest the presence of excess capacity.

The different market designs are judged qualitatively by the degree to which service interruptions are avoided and by the stability of annual average generator revenues under the different conditions.

A.4 Model results

A.4.1 Base case

For the base case, a demand growth of 2.5% per year was chosen because this was the average growth rate between 1977 and 1997 (Sep, 1987; Sep and EnergieNed, 1999).[76] The demand estimation error was set at –1%, which means that investors' demand forecasts are based upon a growth rate of 1.5%, whereas the real growth rate is 2.5%. The sensitivity of the model to these assumptions (except the price of the interruptible contracts) will be evaluated below. Table A.4 sums up the base case parameters. A volume of interruptible contracts of 500 MW was chosen at a price of 2500 €/MWh, as was explained above.

[76] The changes brought about by liberalization were associated with changes in the way data were collected, as a result of which there is a discontinuity in the reported electricity production and consumption data between 1997 and 1998. Therefore a data series ending in 1997 was used to estimate the long-term average growth rate.

Table A.4: Base case settings

Variable	Name	Value	Units
Demand growth	g	2.5	%
Error in projecting demand	e	-1.0	% points
Volume of interruptible contracts	V_{ic}	500	MW

The results for an energy-only market are shown in Figure A.5. The X-axis shows the modeled years, which run from 2004 through 2030. The left Y-axis shows demand and generation capacity, both total capacity and the volume of new capacity in each year. The second Y-axis shows generator revenues.

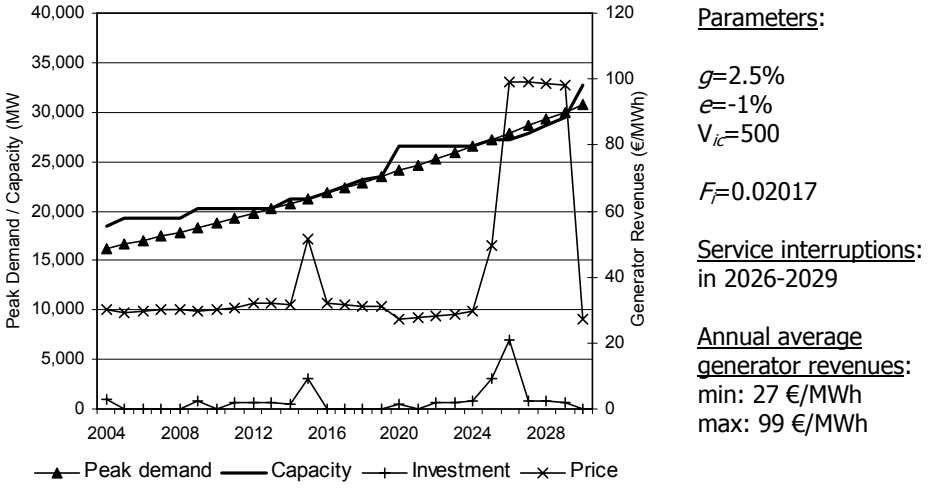


Figure A.5: Base case, energy-only market

In the base case scenario, a period of near shortages leads to a price spike in 2015 that leads to more capacity in 2020.[77] A period with sufficient capacity follows but this is dissipated by demand growth and a new, greater shortage develops with service interruptions. This leads to a strong investment wave, which is followed by a period of overcapacity. During the first shortage in 2015, prices are limited because the interruptible contracts maintain the energy balance. As a result, the highest system prices are set by the interruptible contracts: 2500 €/MWh. The second period of shortages is worse, with black-outs in the years 2023-2026. Consequently, the peak prices rise to the average value of lost load (8600 €/MWh) during these years.

[77] An initial investment of 800 MW is modeled for 2005, which is the Shell/Intergen project that is due to come on-line then (EnergieManagement, 2003).

The fact that generator revenues appear around certain levels (30 €/MWh, 50 €/MWh, 100 €/MWh) is due to the discrete nature of the load-duration curve. When a shortage begins to develop, there is no gradual transition: the top 50 hours of the load-duration curve either have a price equal to the marginal cost of production, or a equal to the cost of interruptible contracts, or equal to the average value of lost load. The relatively large size of this section of the load-duration curve means that as soon as its price is no longer set by the marginal cost of generation, the entire annual average price increases significantly. With more subtle data, the price development in the model would be smoother. The fundamental characteristics would not be different, however.

A.4.2 Sensitivity analysis of the base case parameter settings

To show the sensitivity of the results for an energy-only market to the growth rate g, the demand estimation error e and the volume of interruptible contracts V_{ic}, a number of runs is shown in which these parameters are varied.

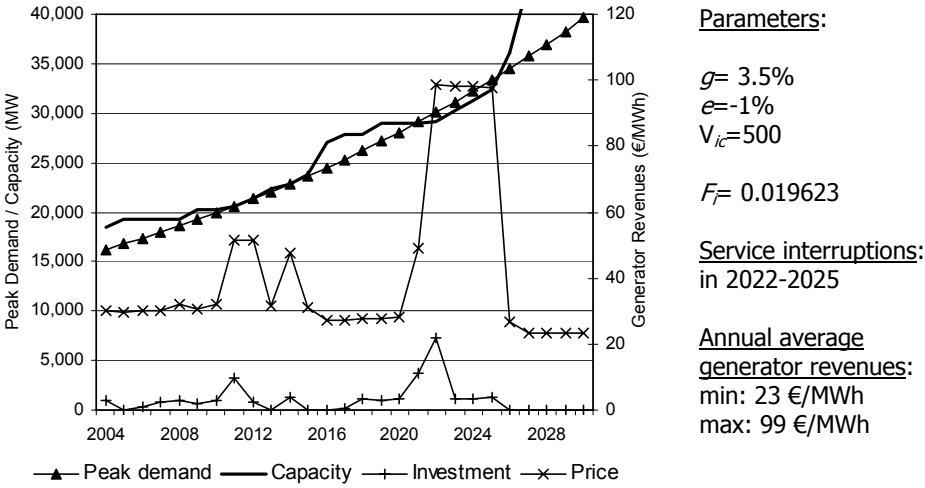
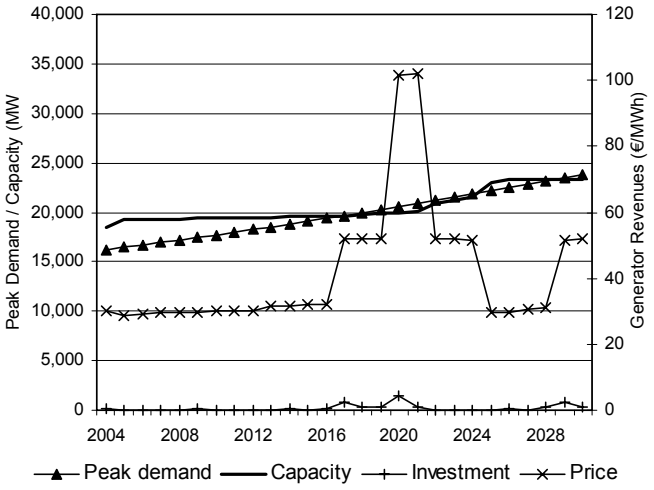


Figure A.6: Energy-only market, high growth rate

The growth rate of demand

Figure A.6 shows the impact of a higher growth rate. A growth rate of 3.5% leads to similar investment cycles as a growth rate of 2.5%. Not surprisingly, the frequency of the shortages increases with the growth rate. At a low growth rate (Figure A.7), investment cycles still appear. The reason is that the price response of investors was modeled to be slower by lowering F_i. This represents the investors' reaction to a low average growth rate; if their reaction to prices was not limited compared to the base case, they would produce more generation capacity, as a result of which the long-run average price would fall below the long-run marginal cost.



Parameters:

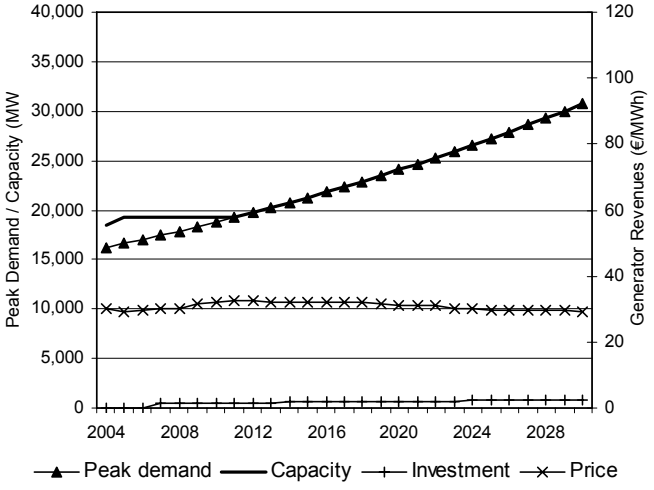
g= 1.5%
e=-1%
V_{ic}=500

F_i= 0.003907414

Service interruptions: in 2020, 2021 and 2030

Annual average generator revenues:
min: 29 €/MWh
max: 102 €/MWh

Figure A.7: Energy-only market, low growth rate



Parameters:

g=2.5%
e=0%
V_{ic}=500

F_i=0

Service interruptions: in 2020, 2021 and 2030

Annual average generator revenues:
min: 29 €/MWh
max: 32 €/MWh
average: 31 €/MWh

Figure A.8: Energy-only market, no demand estimation error

Errors in estimating future demand

Next, the impact of the demand estimation error e will be evaluated. A negative e means that investors do not invest sufficiently to meet the entire growth of demand. Setting e to zero leads to an optimal forecasting of demand, and therefore to a perfectly stable electricity supply, as Figure A.8 shows. Available generation capacity precisely equals demand in this run, due to the perfect foresight of the investors.

This is an unrealistic outcome, however, as the long-run average price is below the long-run marginal cost. In this run, F_i equal zero, which means that generating companies do not react at all to prices and investment is purely driven by their forecast of demand. As a result, it is not possible in the model to obtain an average price of 42 €/MWh without some underestimation of demand.

One would expect that under these conditions generating companies would reduce their investment rate. In practice – or if a continuous load-duration curve had been used – it would be expected that the very tip of the load-duration curve would not be served, so each year there would be shortages equal to the socially optimal duration of outages. These would create some peak revenues, which should be sufficient to increase the average price to the long-run marginal cost of generation.

With the rough load-duration curve that is used in the model, there either are shortages well in excess of the optimal duration of load shedding, and hence with annual average prices in excess of the long-run marginal cost, or there are no shortages at all, like in this run. As a result, a situation with perfect foresight is not stable in this model. Using a more refined (continuous) load-duration curve would solve this problem.

It is more realistic to model at least a small underestimation of demand, so the generating companies can recover their costs. This is shown in Figure A.9. The volume of available generation capacity hovers around peak demand. Limited periods of scarcity develop, during which the interruptible contract volume is sufficient to prevent service interruptions. The reserve margin is nearly nil, however, so the system is vulnerable to changes in demand. In reality, the unpredictability of the growth rate of demand would cause frequent shortages.

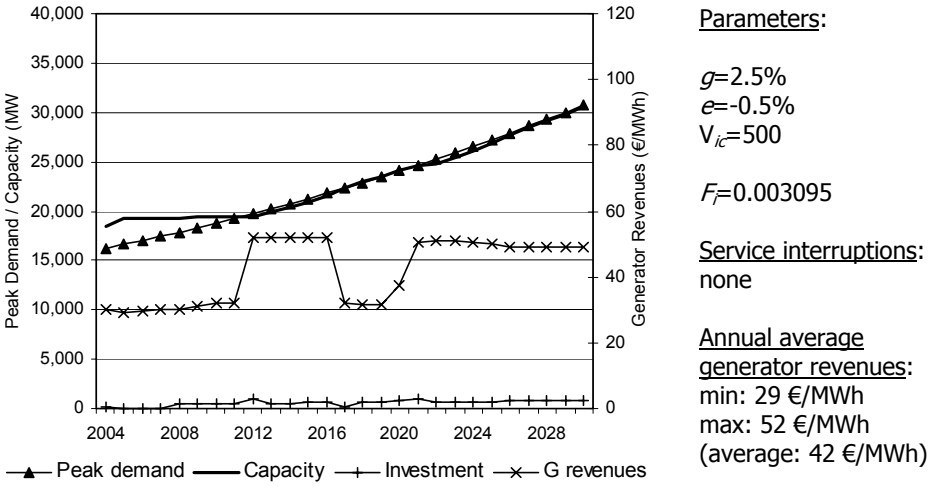
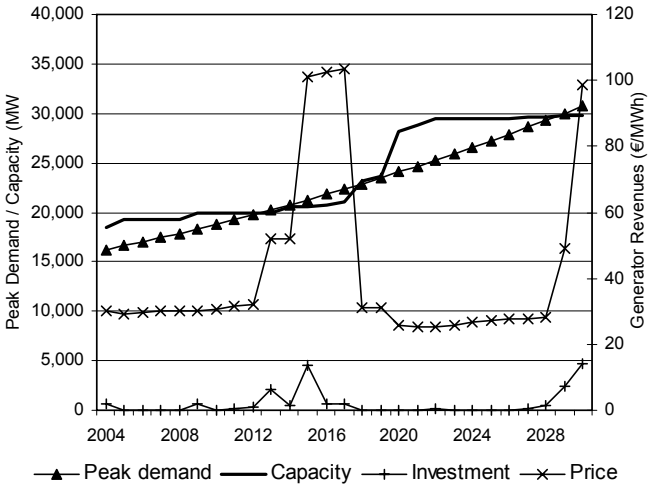


Figure A.9: Energy-only market, small demand estimation error

Figure A.10 shows the effects of a higher demand estimation error *e*. Not surprisingly, the magnitude of the investment cycles has increased. The frequency remains fairly constant throughout most runs, at about one cycle per 12 years.



Parameters:

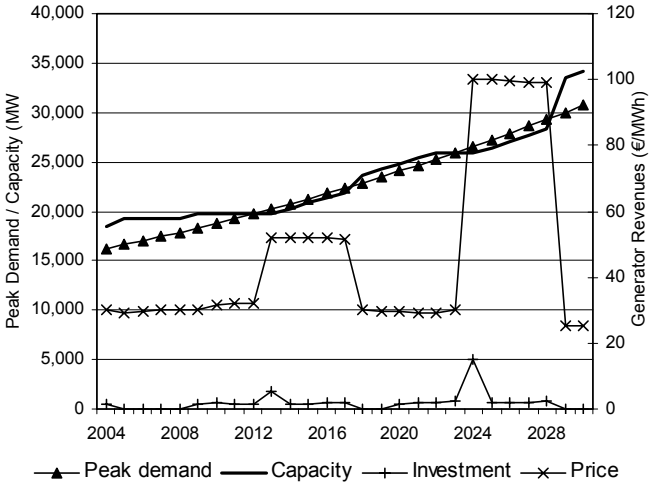
g=2.5%
e=-2%
V_{ic}=500

F_i= 0.01289

Service interruptions:
2015-2017 and 2030

Annual average generator revenues:
min: 25 €/MWh
max: 103 €/MWh

Figure A.10: Energy-only market, high demand estimation error



Parameters:

g=2.5%
e=-1%
V_{ic}=500

F_i=0.01

Service interruptions:
2024-2028

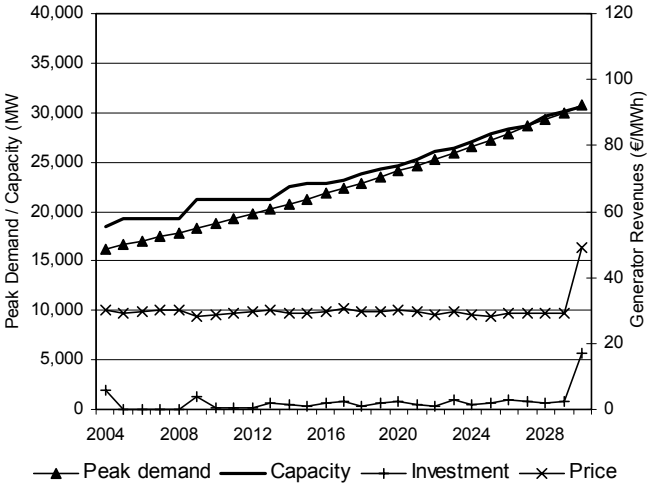
Annual average generator revenues:
min: 25 €/MWh
max: 100 €/MWh
average: 47 €/MWh

Figure A.11: Energy-only market, low investment scaling factor

The investment factor

The investment scaling factor F_i determines how strong an investment reaction to price spikes is modeled. The impact of changes to F_i will be shown by running the base case twice, once with F_i at twice the base case level and once at half the base case level. Figure A.11 shows the effect of a small F_i. A more limited investment response leads to higher average prices. This would be possible if new entry to the generation market were restricted.

Figure A.12 presents the effects of a higher investment factor. Comparing Figure A.11 to Figure A.12, it appears that a higher investment factor F_i (a stronger reaction to price signals) would dampen investment cycles. However, investors do not recover their costs in this case, so this is not a realistic scenario. The system is probably also not as stable as it appears: in case of a demand shock (see Section A.4.6) or, more in general, in case of fluctuations in the growth rate of demand, it may lead to an investment overreaction. In the run of Figure A.12, a strong investment reaction to a limited shortage is seen towards the end of the modeled period.



Parameters:

g=2.5%
e=-1%
V_{lc}=500

F_i=0.04

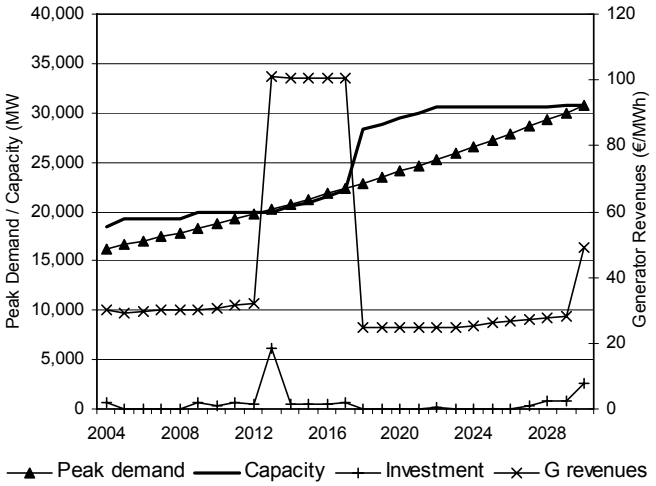
Service interruptions:
none

Annual average generator revenues:
min: 28 €/MWh
max: 49 €/MWh
average: 30 €/MWh

Figure A.12: Energy-only market, high investment scaling factor

Interruptible contracts

Interruptible contracts represent the only demand price-elasticity in the model. The larger their volume, the better the market functions. They provide a buffer between periods of low prices and shortages: when supply is tight, the interruptible contracts are called upon, which reduces demand and lets electricity prices rise substantially. The high prices lead to new investment. The larger the volume of interruptible contracts and the lower the growth rate of demand, the more likely it is that the shortage will not lead to outages before new capacity is available.



Parameters:

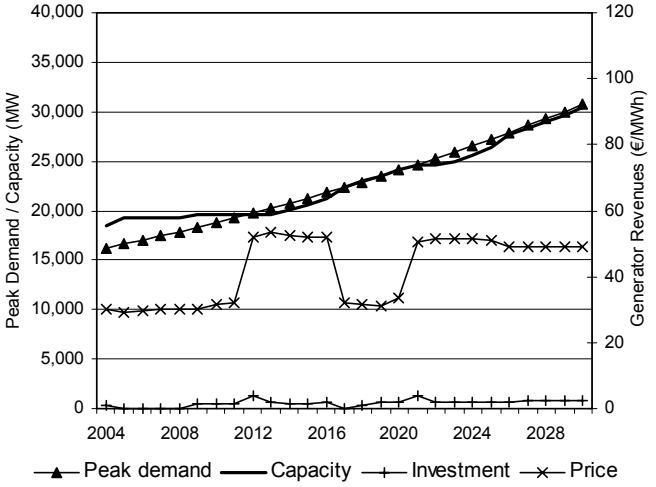
g=2.5%
e=-1.0%
V_{ic}=0

F_i= 0.01324

Service interruptions:
2013-2017

Annual average generator revenues:
min: 25 €/MWh
max: 101 €/MWh

Figure A.13: Energy-only market, no interruptible contracts



Parameters:

g=2.5%
e=-1%
V_{ic}=1000

F_i= 0.006074

Service interruptions:
none

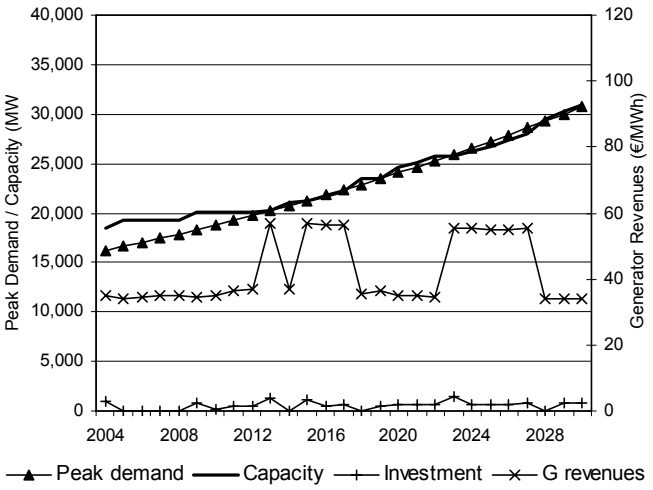
Annual average generator revenues:
min: 29 €/MWh
max: 53 €/MWh

Figure A.14: Energy-only market, high volume of interruptible contracts

Figure A.13 shows that without interruptible contracts, a capacity shortage immediately leads to a period of high price spikes, followed by a prolonged period of outages. Figure A.14 shows that greater demand price-elasticity, in the form of a greater volume of interruptible contracts, dampens the investment cycles. Annual generator revenues still oscillate between about 29 €/MW and 53 €/MW but there are no outages. (At a higher growth rate g or if investors are more risk-averse – e is more negative – outages do occur, however.)

A.4.3 Capacity payments

One of the simplest capacity mechanisms is to provide payments to generating companies for the generation capacity that they keep available. The idea is that if a market would not invest sufficiently in generation capacity, a subsidy would shift the investment equilibrium to a higher level. This is easily modeled because the effect of capacity payments is to lower the fixed costs of generation, so the model can be run with lower fixed costs.



Parameters:

g=2.5%
e=-1%
V_{ic}=500

F_i=0.005836

Service interruptions:
in 2027

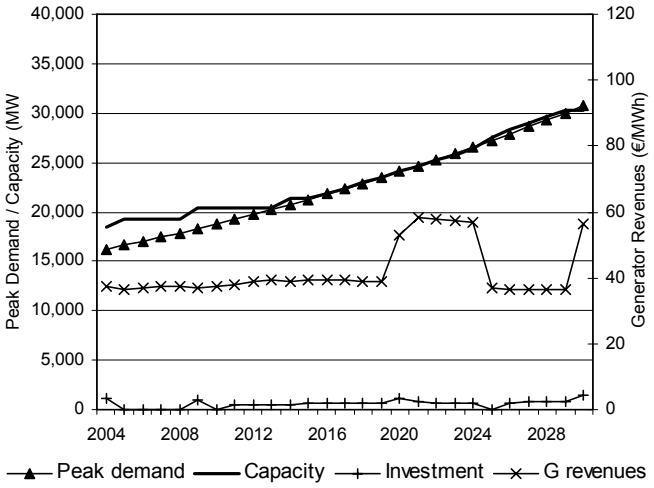
Annual average generator revenues:
min: 34 €/MWh
max: 57 €/MWh

Figure A.15: Capacity payments: the effect of a 50% subsidy on generation capacity

The problem is how to establish the correct level of payments. Figure A.15 shows the impact of a 50% subsidy on the fixed costs of all generation capacity, paid throughout the life of each plant per unit of capacity that is available. The effect of the payments is that new generation becomes attractive at lower prices, so investors respond sooner to price spikes. As a result, investment cycles are dampened but shortages are not prevented entirely. Figure A.16 shows the effects of an increase of the capacity payments to a 75% subsidy. The results are not substantially different from a 50% subsidy. Now there are no shortages but the resulting capacity margin is so small that the system would not be robust against fluctuations in demand.

A.4.4 Operating reserves pricing

Operating reserves pricing is intended to dampen the investment cycle. Investment still is cyclical but the investment signal develops before actual shortages develop. More stable prices and fewer outages result. The effects of operating reserves pricing depends strongly upon the volume of operating reserves and their price, as will be shown below.



Parameters:

g=2.5%
e=-1%
V_{ic}=500

F_i= 0.005249369

Service interruptions:
none

Annual average generator revenues:
min: 36 €/MWh
max: 58 €/MWh

Figure A.16: Capacity payments: the effect of a 75% subsidy on generation capacity

Base case

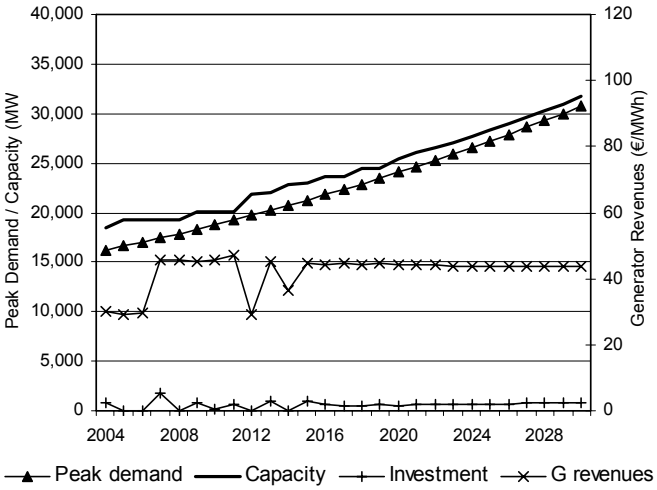
The base case parameters of the operating reserve will be set about equal to those found in Example 6.2 on page 118: an operating reserves volume K_{or} of 2000 MW and a maximum purchasing price P_{or} equal to 2000 €/MWh. These parameters were calculated to provide generating companies with revenues equal to the long-run marginal cost at an optimal average duration of service interruptions, given a certain average value of lost load. The model results of a market with operating reserves pricing with these parameters are shown in Figure A.17.

Due to the static under-estimation of demand growth, the operating reserves are needed to provide a part of peak demand each year. This allows generator revenues to exceed the marginal cost of generation, so a constant investment signal develops.

The model shows no outages at all but again this may be an artifact of the load-duration curve. In an optimally dimensioned system there are some outages, as was argued in Section 5.4.2. Section A.3.4 explained that the model does not show shortages that last only a few hours per year because the top section of the load-duration curve is an average of the 50 highest peak demand hours of the year.

In Figure A.17 the investment cycle eventually disappears. Apparently, a well-designed operating reserves pricing system can provide a stable investment incentive – at least in the perfectly stable conditions of this model. Even in the first investment cycles, the price swings are smaller than in an energy-only market and periods of outages are avoided. If a fixed P_{or} were replaced by a sloping willingness-to-pay curve, as Stoft (2002) suggests, the cycles would probably be reduced sooner. The simplified model that is used here

induces its own volatility by setting the price of the reserve equal to the marginal cost of maintaining generation capacity stand-by when the capacity margin exceeds the size of the reserve, and equal to P_{or} otherwise.



Parameters:

g=2.5%
e=-1%
V_{ic}=500

F_i= 0.01812

K_{or}=2000 MW
P_{or}=2000 €/MWh

Annual average prices:
min: 29 €/MWh
max: 47 €/MWh

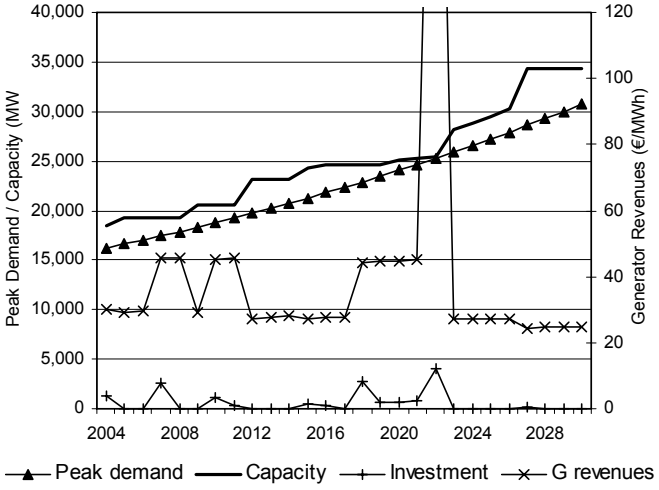
Figure A.17: Operating reserves market, base case

Interestingly, the model has multiple solutions where the long-run average revenues of generators are equal to the long-run marginal cost. See Figure A.18: like in Figure A.17, average generator revenues are 42 €/MWh. However, prices are much less stable: annual average prices range up to 281 €/MWh in 2022, when service interruptions are just avoided. This poses the question which solution to chose. As a pragmatic solution, the value for F_i is chosen that is closest to the base case value.

Sensitivity to the operating reserve volume and price

Now the question will be addressed how operating reserves pricing functions under different circumstances and with different settings. Figure A.19 shows the effect of a smaller operating reserve with a higher price, an option which also was calculated in Example 6.2 on page 118. Due to the smaller volume of operating reserves, the system now begins to resemble an energy-only market, with investment cycles which eventually lead to shortages.

A substantially larger operating reserve than the base case of 2000 MW is not easily feasible, as in that case the reserve requirements are not met at the beginning of the modeling period. A transition period would be required.



Parameters:

g=2.5%
e=-1%
V_{ic}=500

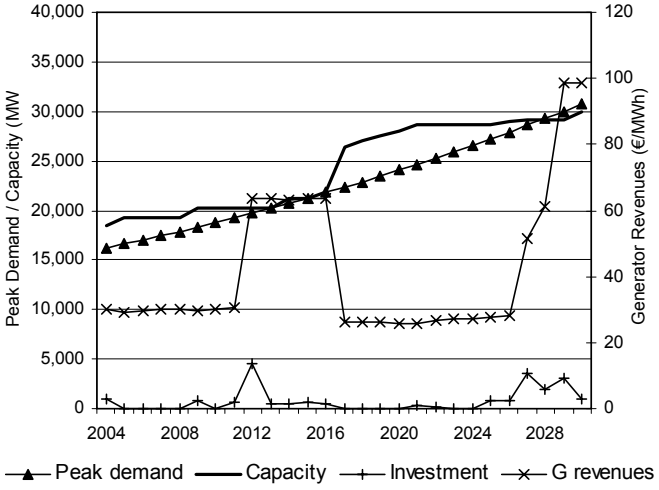
F_i= 0.026158742

K_{or}=2000 MW
P_{or}=2000 €/MWh

Service interruptions:
none

Annual average prices:
min: 24 €/MWh
max: 281 €/MWh

Figure A.18: Operating reserves market, equilibrium with shortages



Parameters:

g=2.5%
e=-1%
V_{ic}=500

F_i= 0.019841736

K_{or}=1000 MW
P_{or}=4000 €/MWh

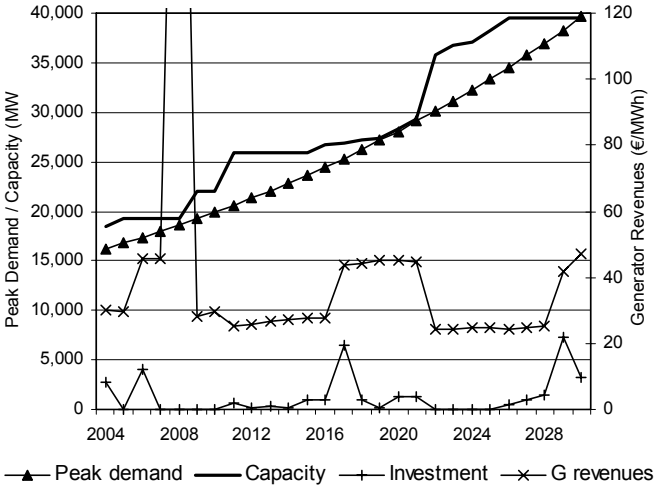
Service interruptions:
in 2029 and 2030

Annual average prices:
min: 26 €/MWh
max: 99 €/MWh

Figure A.19: Operating reserves pricing, smaller reserve with a higher price

High growth rate of demand

Transition problems also occur in the case of a higher growth rate of demand, as is shown in Figure A.20. The most striking feature of this run is the extremely high annual average price in 2008: 281 €/MWh. There is no physical shortage during this year but the operating reserves price determines the market price for about 10% of the time during that year. This is a transition effect because after the introduction of operating reserves pricing in 2004 there was not enough time to create more generation capacity before 2009. The conclusion presents itself that in a market with limited spare capacity, the introduction of a large volume of operating reserves immediately leads to a situation in which the demand for operating reserves is not satisfied. If the system had already been in place, it would have generated an earlier investment signal. Generation capacity would already have been in the pipeline in 2004, which would have resulted in more generation capacity before 2008.



Parameters:

g=3.5%
e=-1%
V_{ic}=500

F_j= 0.055194294

K_{or}=2000 MW
P_{or}=2000 €/MWh

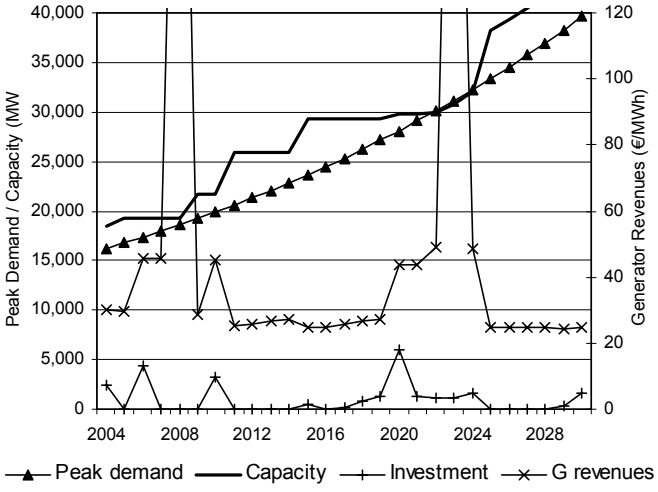
Service interruptions:
none

Annual average prices:
min: 24 €/MWh
max: 47 €/MWh
(281 €/MWh)

Figure A.20: Operating reserves market, high growth rate, low long-run average revenues of generators

An important lesson is that if the existing capacity margin is smaller than the desired size of the operating reserves pricing, there is a need for a transition regime that gradually increases the investment signal. Otherwise, operating reserves pricing will lead to average prices far above the long-run marginal cost of generation capacity (even if peak prices are limited to the operating reserves price) until new generation capacity has been constructed. Creating a transition phase may be difficult. The size of, and/or the price paid for the operating reserves must be increased gradually to avoid unduly large income transfers from the consumers to the generators but the investment signal must be strong enough so the capacity margin increases to the level demanded by the system operator.

If the assumption is made that there is a transition period, the price spike in 2008 should not be included in the analysis. After 2008, the highest annual average price is 47 €/MWh. However, the long-run average electricity price between 2009 and 2030 is only 32 €/MWh, which is well below the long-run marginal cost of generation. What happens if the investment signal is re-calibrated so the long-run average revenues of the generating companies after 2009 are equal 42 €/MWh? The results are presented in Figure A.21. Including the transition phase, the long-run average revenues of the generating companies are 50 €/MWh. After the transition phase, the generator revenues equal the long-run marginal cost of generation on average but now a second price spike occurs, equally high as the transition price spike in Figure A.20. Again, there are no physical shortages; the high annual average price is caused by the fact that due to the slim capacity margin, the operating reserves price determines the market price a significant part of the time that year. Apparently, operating reserves pricing is not stable in the presence of such a high growth rate.



Parameters:

g=3.5%
e=-1%
V_{ic}=500

F_i= 0.04955

K_{or}=2000 MW
P_{or}=2000 €/MWh

Service interruptions:
none

Annual average prices:
min: 24 €/MWh
max: 284 €/MWh
average: 50 €/MWh

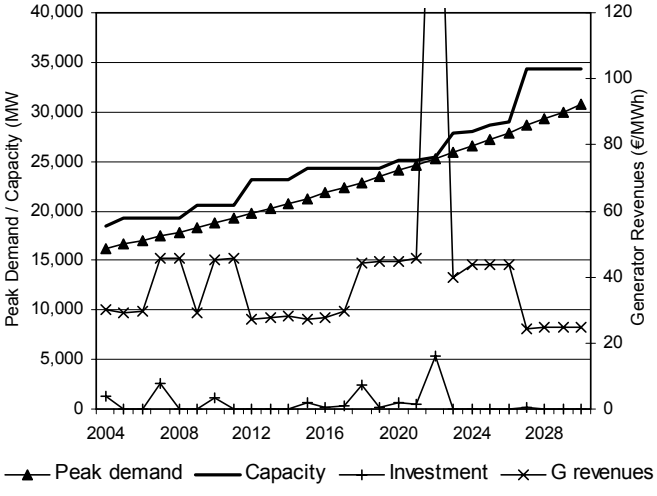
Figure A.21: Operating reserves market, high growth rate, long-run average revenues of generators equal to long-run marginal cost

By comparing Figure A.20 and Figure A.21, the impact of small changes to F_i becomes apparent: a difference of only about 10% can lead to significantly higher price spikes. Apparently, the stability of the system remains highly sensitive to investors' reaction to electricity prices.

Higher demand forecasting error

Next the robustness of operating reserves pricing with respect to a chronic under-estimation of future demand is evaluated. An extreme case will be considered, in which the demand growth rate is 2.5% (as in the base case) and the demand estimation error e=-

2.0%. This means that investors forecast a growth of demand of only 0.5% per year. The results are shown in Figure A.22.



Parameters:

g=2.5%
e=-2.0%
V_{ic}=500

F_j= 0.026158742

K_{or}=2000 MW
P_{or}=2000 €/MWh

Service interruptions:
none

Annual average prices:
min: 24 €/MWh
max: 216 €/MWh

Figure A.22: Operating reserves market, large error in demand forecasts

Considering the extreme underestimation of demand, the system appears quite stable. Outages are avoided and only one year with high prices develops. However, this price spike develops only 18 years into the model, so the question is whether investors would have kept up the investment rate during this time. If the investment scaling factor is corrected, like was done before, to create higher average generator revenues during the first period, the investment signal is reduced. While this more cautious investment behavior raises the average price in the begin period, now a severe shortage develops from 2026 on, resulting in outages and years of high price spikes before a significant wave of new generation capacity becomes available. Again, the conclusion is that the system may become instable: periodic high price spikes keep occurring.

It should be kept in mind, however, that this is an extreme case: if investors would only anticipate a demand growth rate of 1.0% instead of 0.5%, there would be no price spikes or shortages and the results become similar to the operating reserves pricing base case, except that the reserve margin becomes quite slim. The latter is a vulnerability in the reality of changing demand growth rates, which means that investment cycles may yet develop.

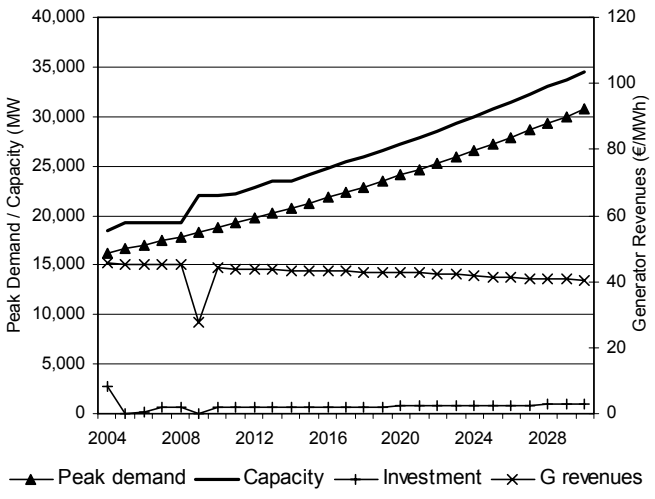
In conclusion, operating reserves pricing dampens the investment cycle but may not be stable under all conditions. While the probability of outages is reduced, high annual average prices may still develop from time to time. This apparent lack of stability is troubling because the model environment is perfectly stable. Operating reserves pricing will probably perform less well in the reality of varying growth rates of peak demand. Section A.4.6 analyzes the impact of a single demand shock.

A.4.5 Capacity requirements

Of the capacity mechanisms that have been modeled here, a system of capacity requirements has the most profound impact upon the dynamic development of the electricity system. By creating an explicit demand for capacity, a stronger signal is provided to invest, even when current generation capacity is sufficient to meet demand. The resulting capacity margin may appear excessive in the static conditions of this model but serves as a buffer for unanticipated demand or supply shocks.

Essentially, a system with capacity requirements takes the responsibility for planning generation capacity away from the market and places it with a system planner. It may be questioned whether he would be better at this than market parties, but for the reasons mentioned in Chapter 5, he may be better able to represent the consumers' interests, which is to err on the side of over-investment, if erring is inevitable. This is not fundamentally different from operating reserves pricing, as there a central planner also needs to establish the optimal volume of generation capacity. (See Section 6.4.) However, in that case a price incentive is used to obtain the desired volume of generation capacity, rather than more direct regulation.

The main parameters of a system of capacity requirements are the reserve requirement r, which is 17% of current demand in the base case, and the penalty that the load-serving entities pay if they are short of their capacity requirements, which is set at 87,640 €/MW per year (equal to the fixed costs of a new plant). This does not play a role in the model, however, as it is assumed that the penalty is high enough so load-serving entities try to meet their capacity requirements.



Parameters:

g=2.5%
e=-1.0%
V_{ic}=500
r=17%

F_i=0.0155

Service interruptions:
none

Annual average generator revenues:
min: 28 €/MWh
max: 46 €/MWh

Figure A.23: Capacity requirement, base case

Figure A.23 shows the base case results for a market with a capacity requirement. Generator revenues are composed of the price that consumers pay for electricity plus the price of the capacity credits. As long as there is ample capacity, the electricity price in the model is equal to the marginal cost of generation, which is around 25 €/MWh. The capacity credit price is equal to the penalty price for load-serving entities that are deficient, unless the availability of capacity credits exceeds the capacity requirement. In the static conditions of this model run, the capacity market always is slightly constrained except in 2009, when a slight overreaction to the scarcity of capacity credits in 2004 leads to a temporary glut of capacity credits. In this year the capacity credit price drops to nearly zero.

Again, average generator revenues are calibrated to be equal to the long-run marginal cost of generation. The lost price spike income is replaced with revenues from the capacity market. The higher cost of more generation capacity is offset by the fact that the new capacity has lower operating costs, so the electricity price drops. In this model, these effects happen to balance each other roughly out. In practice, this is of course not necessarily the case, so it remains important to keep the capacity margin as near the optimum as is possible.

Sensitivity of the design of the capacity requirement

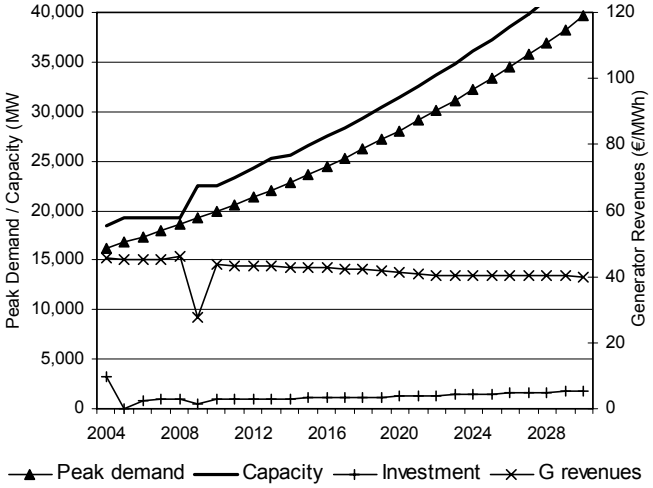
With the base case parameters and a fixed demand growth rate, a higher reserve requirement r simply results in a higher reserve. The size of the reserve only is important in the case of supply or demand shocks. Therefore its optimal size cannot be deduced from the base case. The main investment signal is provided by the investors' forecast of the capacity requirement, even if the model is constructed so they strongly underestimate demand growth. Response to price spikes only plays a minor role. This is true in practice as well, witness the fact that the PJM system has a price cap of 1000 $/MWh in the energy market.

The limited role of the price signal for investment also means that the impact of the investment scaling factor F_i is small. It will continue to be adjusted each run, so the long-run generator revenues equal the long-run marginal cost. Due to its limited impact, the variations in F_i will be large. However, if F_i is kept the same as in the base case, the model results do not differ significantly.

High growth rate and high demand estimation error

Figure A.24 shows that a system with a capacity requirement is quite robust with respect to a high growth rate of demand. As in the case of operating reserves pricing, a high growth rate may give rise to a transition effect, although in this case it is much smaller. Nevertheless, capacity requirements may need to be phased in if the existing capacity margin is substantially smaller than the desired margin.

Figure A.25 shows a situation with the same high growth rate of 3.5% per year and a stronger under-estimation of demand growth. The demand estimation error is set at –2.0%, which means that the forecast growth is 1.5% per year. The system still functions satisfactorily.



Parameters:

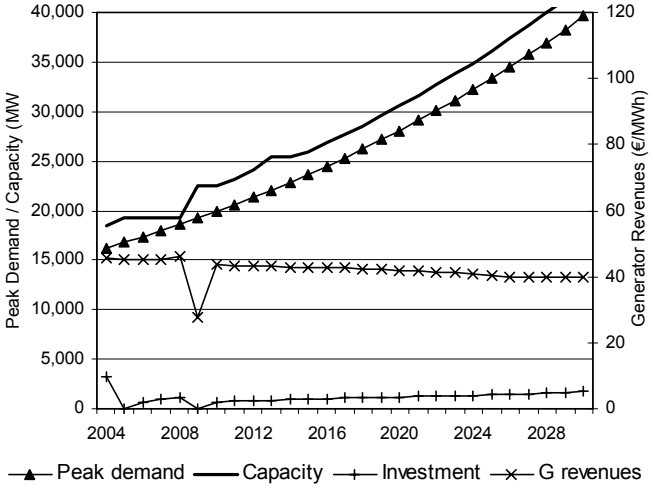
g=3.5%
e=-1.0%
V_{ic}=500
r=17%

F_i= 0.005388

Service interruptions:
none

Annual average generator revenues:
min: 28 €/MWh
max: 46 €/MWh

Figure A.24: Capacity requirement, high growth rate



Parameters:

g=3.5%
e=-2.0%
V_{ic}=500
r=17%

F_i= 0.02633

Service interruptions:
none

Annual average generator revenues:
min: 28 €/MWh
max: 46 €/MWh

Figure A.25: Capacity requirement, high growth rate and large error in demand forecasts

It may be concluded that capacity requirements are the most successful capacity mechanism in stabilizing investment. The fact that capacity requirements provide a clearer indication of the future demand for generation capacity, even if the demand growth rate is underestimated, probably contributes indeed to a more stable investment signal. The main question is how this system performs under a less regular development

of demand.

A.4.6 Demand shock

An interesting test of the capacity mechanisms is their robustness against a demand shock. A scenario in which demand increases by an additional 15% over seven years (2015-2021) is used. This is the equivalent of a demand growth rate that is 2 percentage points higher during that period. This scenario could develop for instance if the Netherlands implements a capacity mechanism while its neighboring countries do not: then it is conceivable that in the long term a regional shortage develops during peak demand periods, so the Netherlands can no longer count on cheap imports (which currently provide in about 15% of demand).

A sudden decrease in the availability of imports is most easily modeled as a corresponding increase in demand. Therefore the model runs in this section may also be interpreted as having demand shocks from other causes; the point is to investigate the robustness of the different capacity mechanisms. It will be assumed that the generating companies did not expect the demand shock; therefore the investment factor F_i will be kept the same as in the corresponding scenarios without a demand shock.

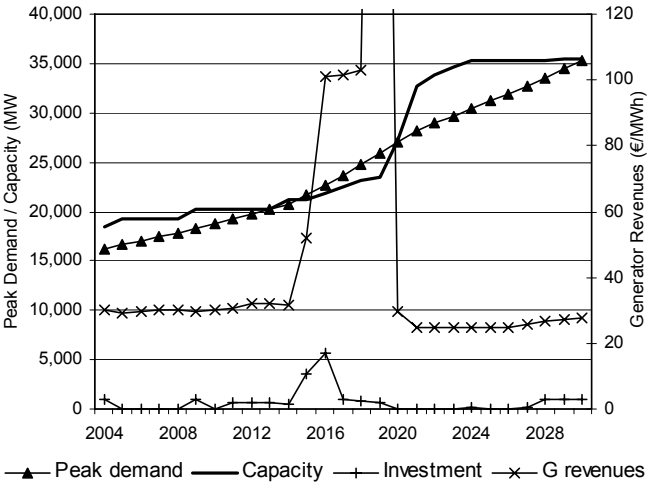
Figure A.26 shows the impact of the demand shock upon the base case scenario for an energy-only market. The demand shock leads to a serious crisis, with physical shortages during the first five years of the demand shock. For three years in a row, 2019-2021, the annual average price rises to nearly 400 €/MWh. An overreaction takes place, followed by a period of extremely low prices, as a result of which another shortage is imminent towards the end of the modeled period. (Note that the long-run average generation revenues no longer equal the long-run marginal cost of generation because the investment factor F_i has been kept the same as in the corresponding model runs without a demand shock.)

The situation is worse if the regular growth rate of demand g is low, as Figure A.27 demonstrates, even if the demand estimation error e also is small. With these settings an energy-only market would perform well under normal conditions (Figure A.9). Now the period with outages lasts longer, from 2016 through 2022, and prices become excessively high. However, even if the investment factor F_i is recalibrated, so the long-run generator revenues are equal to the long-run marginal cost, a similar investment cycle develops (Figure A.28). The stronger investment reaction reduces the magnitude of the shortage but leads to an excessive investment reaction. In conclusion, an energy-only market is not robust with respect to a significant trend change in the growth rate of demand.

Capacity payments

Applying capacity payments, even a 75% subsidy of the fixed costs, does not stabilize the system in the face of a demand shock, as Figure A.29 shows. In 2019 an extreme shortage develops: during about 1650 hours of the year, supply is inadequate. This causes an extreme price spike, due to which the long-run average electricity price (and hence the generator revenues) is much higher than in the case of an energy-only market. If the

investment signal is modeled as being more price-responsive (with a higher F_i), the effect is more limited but an investment cycle still develops, with several years of shortages followed by significant overinvestment.



Parameters:

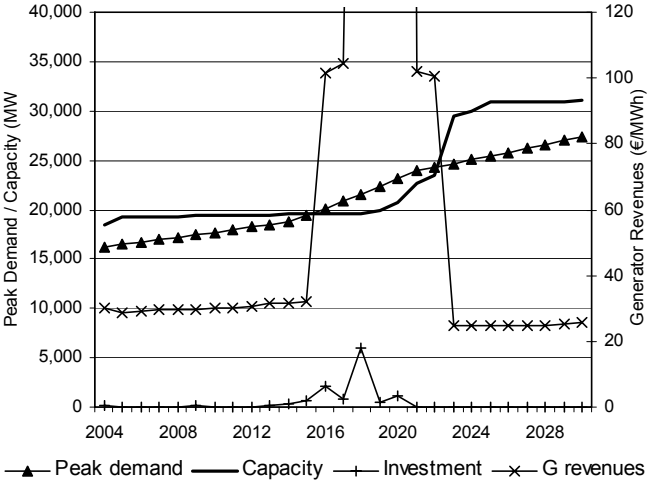
g=2.5%
e=-1.0%
V_{ic}=500

F_i= 0.02017

Service interruptions:
2016-2019

Annual average generator revenues:
min: 25 €/MWh
max: 398 €/MWh
average: 51 €/MWh

Figure A.26: Energy-only market, base case parameters, demand shock of 2 percentage points between 2015 and 2022



Parameters:

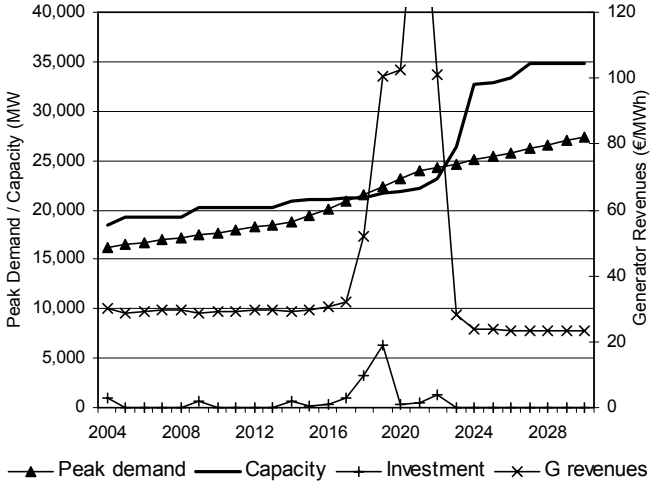
g=1.5%
e=-0.5%
V_{ic}=500

F_i= 0.01983

Service interruptions:
2016-2022

Annual average generator revenues:
min: 25 €/MWh
max: 1148 €/MWh
average: 135 €/MWh

Figure A.27: Energy-only market, low base demand growth rate, small demand estimation error, demand shock of 2 percentage points between 2015 and 2022



Parameters:

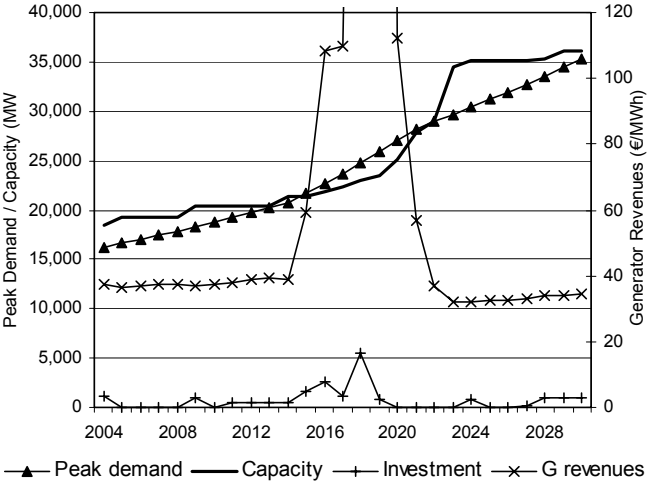
g=1.5%
e=-0.5%
V_{ic}=500

F_i= 0.003907414

Service interruptions:
2019-2022

Annual average generator revenues:
min: 23 €/MWh
max: 171 €/MWh
average: 42 €/MWh

Figure A.28: Energy-only market, low base demand growth rate, small demand estimation error, demand shock of 2 percentage points between 2015 and 2022, recalibrated investment factor



Parameters:

g=2.5%
e=-1.0%
V_{ic}=500

F_i= 0.005249

Service interruptions:
2016-2020

Annual average generator revenues:
min: 32 €/MWh
max: 1137 €/MWh
average: 100 €/MWh

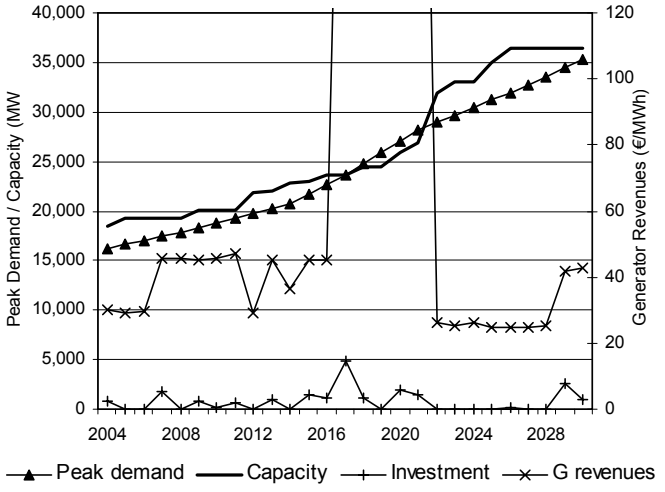
Figure A.29: Capacity payments (75% of fixed costs), base case parameters, demand shock

Operating reserves pricing

Under the base case conditions, operating reserves pricing hardly provides an improvement, as Figure A.30 shows. An investment cycle develops with three years with

service interruptions and the average electricity price is significantly higher than in an energy-only market under the same conditions. (See Figure A.26). In this case, operating reserves pricing does not stabilize investment but it does creates substantial additional income transfers from consumers to producers, well in excess of an energy-only market.

If the investment behavior is modeled to foresee the demand shock, the average generator revenues could be much lower. However, this would require investors to build up substantial overcapacity in expectation of the demand shock. Between the realization of this large capacity margin and the occurrence of the demand shock, competition would force prices to below the long-run marginal cost of generation. Given the many uncertainties in electricity markets, in particular about demand growth several years into the future, it was considered unlikely that investors anticipate trend changes in the growth rate of demand. Therefore they were assumed to behave the same as in the corresponding model runs without a demand shock (F_i was kept the same), as the introduction to this section explained.



Parameters:

g=2.5%
e=-1.0%
V_{ic}=500

F_i= 0.01812

K_{or}=2000 MW
P_{or}=2000 €/MWh

Service interruptions:
2019-2021

Annual average generator revenues:
min: 25 €/MWh
max: 349 €/MWh
average: 88 €/MWh

Figure A.30: Operating reserves pricing, base case scenario + demand shock

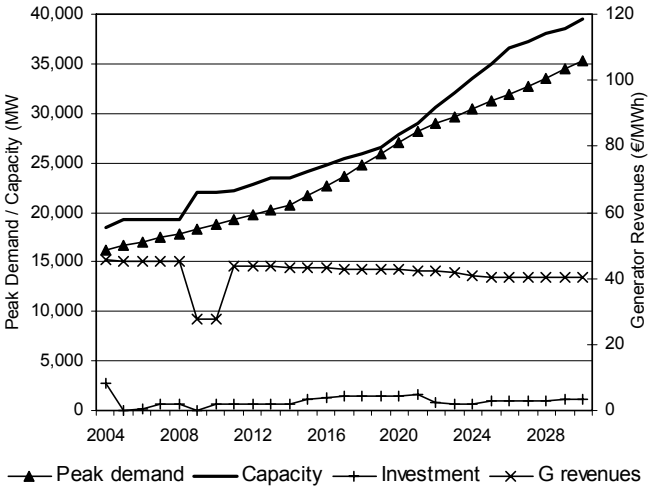
During the years with service interruptions, the average electricity price is about 340 €/MWh. The cause of the higher electricity prices is that when outages occur, the electricity price rises to the average value of lost load in both cases, so then there is no difference between the two market designs. However, during off-peak hours the price drops faster in an energy-only market than in the case of operating reserves pricing. Operating reserves pricing lengthens the price spikes, as just before and after the outages the price is determined by the system operator's willingness to pay (P_{or}). When the capacity margin is slim, the electricity price may be determined by P_{or} much of the time. This mechanism provides an early investment signal and therefore has a stabilizing effect during normal conditions, but *if* it fails to cause timely investment, it aggravates the cost

of price spikes to consumers. The prolonged high prices may lead to an overreaction by investors, such as in the latter part of the run in Figure A.30.

Capacity requirements

Finally, the effects of demand shock upon a system with capacity requirements will be considered. In the system of capacity requirements it is possible to ensure that there is adequate generation capacity within the system to meet demand; if this is the case, the scenario of a sudden reduction of imports would not create a demand shock. However, let us assume there is another reason why demand suddenly increases significantly.

The results are shown in Figure A.31. The model parameters are the same as in Figure A.23. The system performs as desired: the long-run average price is near the long-run marginal cost, prices are stable (annual average prices hardly vary) and there are no shortages. Even the long-run average revenues of the generating companies remain equal to the long-run marginal cost. A system with reliability contracts would probably behave similarly, as the demand for generation capacity would be equally predictable. The conclusion presents itself that the capacity-based systems (capacity requirements and reliability contracts) are the most robust with respect to changes in demand.



Parameters:

g=2.5%
e=-1.0%
V_{lc}=500

F_j=0.01550

Service interruptions:
none

Annual average generator revenues:
min: 28 €/MWh
max: 46 €/MWh
average: 42 €/MWh

Figure A.31: Capacity requirements, base case scenario + demand shock

A.5 Conclusions

A model was developed to evaluate the differences in the long-term dynamic behavior of several capacity mechanisms. In the model of an energy-only market, investment cycles develop in most runs as a consequence of the following assumptions:

- perfect competition (which precludes strategic investment behavior)

- a delay in the availability of new capacity
- imperfect information regarding the future demand for generation capacity.

The model was used to evaluate the robustness of capacity payments, operating reserves pricing and capacity requirements against the development of investment cycles.

The main conclusion is that capacity mechanisms with a regulated volume of generation capacity are more robust than those that use economic incentives for stimulating investment. Of the modeled capacity mechanisms, capacity requirements therefore are the most attractive. Reliability contracts were not modeled but should perform similarly.[78] Even under quite extreme circumstances, such as a reasonably high annual growth rate of 2.5% plus a demand shock of another 2 percentage points growth during seven years, the a capacity requirement of 17% maintains low prices and avoids shortages. This corroborates the finding of Chapter 7 that providing a clear and unambiguous demand signal for capacity is the most effective way to avoid investment cycles.

Capacity payments, if they are large enough, provide some stabilizing force to an energy-only market but are not robust with respect to fluctuations in the growth rate of demand. Operating reserves pricing provides an improvement in some cases but may become instable if investment lags growth by too much, if the growth rate of demand is too high or as a consequence of demand shocks. Operating reserves pricing therefore appears to provide only a limited benefit.

An important difference between capacity requirements and operating reserves pricing is that the former not only reacts better to demand and supply shocks but also prevents some of them. By requiring investment to take place within the system, capacity requirements preclude a scenario in which scarcity suddenly develops due to a drastic reduction of imports. Operating reserves pricing does not provide this option. Conversely, in a system of capacity requirements (at least in the integrated model of PJM), exports can be recalled during a shortage, so a volume of generation capacity equal to demand plus the capacity margin is available to the users in the system. With operating reserves pricing, it is not possible to recall exports.

Another lesson is that implementation of a capacity mechanism may create a difficult transition phase if the existing capacity margin is smaller than the desired margin. The reserve requirements may need to be expanded gradually, without undermining the effectiveness of the investment incentive. It will be much easier to implement a capacity mechanism in a situation in which the capacity margin is larger than intended, so the investment signal can and increase gradually.

A.6 Research recommendations

The main shortcomings of this model are that it is deterministic, whereas the stochastic variations of available generation capacity and demand are crucial factors, and that it

[78] The model does not distinguish between the two, as the two systems mainly differ with respect to the non-competitive behavior of generating companies, which was not modeled.

ignores market power. The analysis of the effects of a demand shock suggest that the robustness of capacity mechanisms to unpredictable fluctuations in the growth rate of demand may vary considerably. Therefore a first improvement of the model would be to let demand grow in a stochastic manner. Monte Carlo simulations could be used for this purpose.

Equally important, but probably difficult, would it be to model oligopolistic behavior. Two separate market power issues exist. The first is the impact of market power upon the equilibrium prices. An approach to this problem has been developed by Day et al. (2002).

The second issue is the question how market power would affect investment behavior. Section 5.6.2 provided a brief introduction to this issue. In general, the way that investment decisions are modeled could be improved. Real options theory could provide a framework for the modeling of investment decisions (Dixit and Pindyck, 1994).

Another omission in the model that should be corrected is the fact that it does not model the decommissioning of plant. During the 'bust' phase of an investment cycle, it is likely that old plants are decommissioned, which would shorten the period with low prices and therefore increase the long-term average price.

Finally, the model results would be refined if better load-duration data would be used and if the equilibrium prices would be calculated for shorter intervals than per year.

Summary

The principle underlying the liberalization of European electricity markets is that the network, as a natural monopoly, continues to require regulation, but generation can be provided by a competitive market that does not require specific regulation. This study challenges the latter assumption. The specific characteristics of the electricity system, such as the difficulty of storing electricity, the long life cycle of generators and the even longer life cycle of electricity networks, and the close relationships between generation and network cause the dynamics of the electricity sector to be different from other markets. Considering the high social and economic costs of disruptions of electricity service, electricity markets should be designed with great care.

The central research question of this study is:

Does the current design of European wholesale electricity markets provide adequate long-term incentives for achieving reliability and economic efficiency, and if not, what are the policy options for intervention?

The research focuses on two aspects of investment in electricity generation capacity in European electricity markets: the quantitative question of whether the market will continue to provide sufficient generation capacity and the qualitative question of how to coordinate investment in generation capacity with the physical capabilities of the electricity network. Chapter 2 introduces the research question, scope, and method. Chapter 3 describes the view of the electricity system that underlies the dissertation.

Generation adequacy

The analysis of the question of generation adequacy begins with a case study of the electricity crisis in California in 2000 and 2001 (Chapter 4). This case study is a reference point for both public policy, due to its high and well-publicized economic impact, and for scientific research, as it provides a complex but fascinating body of evidence. The conclusion of the analysis is that, despite the many idiosyncrasies of the California market design and the many mistakes that were made, the basic factors that caused the crisis were not unique to California:

- For years, investment in generation capacity had been lagging behind demand growth, even though generating companies turned out to be able to make high profits when shortages developed.

- The low price-elasticity of demand, combined with the absence of long-term contracts, allowed generating companies to raise wholesale electricity prices far above their competitive levels by withholding generation capacity from the markets. This contributed significantly to the interruptions of electricity service and to the high prices paid by consumers.
- The crisis was triggered by a reduction of imports from neighboring states with different market structures. In the face of limited supply, these gave their own consumers priority.

Chapter 5 provides an analysis of why competitive energy-only markets, in which the electricity price is the only driver of investment in generation capacity, do not always provide a socially acceptable volume of generation capacity. The low price-elasticity of demand and the difficulty of storing electricity cause electricity prices in most markets to be highly volatile. This causes investment in generation capacity to be risky. In theory, risk-neutral investors should produce a socially optimal volume of generation capacity but regulatory risk and a lack of market transparency may easily cause the investment equilibrium to be lower than the social optimum.

The main risk to consumers is not so much static under-investment as the risk of investment cycles. The long lead time for new generation capacity causes the investment response to price spikes to be delayed by a number of years. Even if the average level of investment were optimal, investment cycles would be costly to consumers, as was demonstrated by the electricity crisis in California.

When the optimal volume of generation capacity (which, of course, changes over time) is cannot be achieved, the interest of consumers is to err in favor of excess generation capacity. The cost of excess generation capacity is limited compared to the costs of shortages. Moreover, the cost of excess generation capacity would at least partially be compensated by a reduction of market power. For these reasons, electricity markets should have a 'capacity mechanism', a system of rules and incentives with the purpose of stabilizing the volume of generation capacity. Currently, most European electricity markets do not have a capacity mechanism.

Chapter 6 presents several existing and proposed capacity mechanisms. Chapter 7 develops a framework for the evaluation of capacity mechanisms and applies it to the capacity mechanisms presented in Chapter 6. The first conclusion is that capacity mechanisms that directly control the volume of generation capacity (capacity requirements, reliability contracts and capacity subscriptions) are preferable to ones that work indirectly through the use of price incentives.

Given the international nature of European electricity markets, the best solution would be for interconnected electricity systems to implement a capacity mechanism jointly. This is important not only for the sake of avoiding distortions of international trade but also because it is much more difficult to develop an effective capacity mechanism in a decentralized system if neighboring systems do not have a similar capacity mechanism. However, it is questionable whether it will be possible to achieve this goal in time to avoid a first downswing in the investment cycle.

Individual European countries may choose to implement a capacity mechanism unilaterally while they wait for a regional solution to be developed. Unfortunately, none of the capacity mechanisms that are described in Chapter 6 are effective in an open, decentralized system. During a regional shortage, the generation capacity that was partly paid for through the capacity mechanism might still be used to export electricity, as a result of which the consumers in the system with the capacity mechanism would not experience lower prices or better reliability than without it.

Chapter 8 explores some innovative options for the specific case, typical of Europe, of implementation in an open, decentralized system. Two variants of reliability contracts are the most interesting medium-term solution. The centralized version has the disadvantages of being complex and relying on a auctions that might be susceptible to the exercise of market power. A bilateral version may not be completely effective in the presence of vertical integration of generation and retail companies.

Coordination

The second aspect of investment in generation capacity to be investigated is the issue of how to coordinate the generation market with the network (Chapters 9 and 10). To create a level playing field in the generation market, it is essential that generating companies are 'unbundled' from network companies, so network companies have no economic interests in the generation market. Physically, however, there is a need to coordinate operation and development of generation and the networks with respect to:

- operational behavior of generating companies,
- the management of reactive power by generating companies, and
- locational decision by generating companies.

In a liberalized market, the system operator does not have the authority to plan generation operation and investment. Therefore the ideal way to meet the need for coordination would be to create efficient economic incentives for generating companies. This implies that the goal for network charges should not only be to recover the cost of the network but also to provide generators with incentives to use the network efficiently. One possible solution that provides efficient incentives to generators is the system of locational marginal pricing of access to the electricity system, as is used in several electricity systems in the USA and New Zealand.

Europe has chosen for a different market structure, at least partly out of necessity: the institutional requirements for locational marginal pricing appear to be too high for European states to meet in the short term, due to their diverging legal, institutional and market structures. This research project has investigated the consequences of the European choice for fixed network charges. In addition, the goal was to stimulate the market through a simple and transparent system of network access. However, the presence of continually changing network externalities means that fixed network charges cannot provide efficient economic incentives. Thus they hamper the long-term coordination of the development of networks and generation stock.

Policy options for improving the operational and long-term incentives to generating companies in systems with fixed system access tariffs include the use of congestion

management methods, variations in connection charges, payments for ancillary services and permits. A combination of these incentives will result in a pragmatic mix rather than a theoretically consistent system. The mix of incentives will need to be recalibrated from time to time to obtain the desired effect. Thus, the paradoxical situation may develop of a complicated, intransparent combination of *ad hoc* fixes to mitigate the negative side-effects of fixed network access tariffs that were established to provide simplicity and clarity.

Considering the lack of data on this issue, the development of markets should be monitored with respect to the efficiency of decisions by generating companies regarding investment in and retirement of generators. These decisions should be compared to what would be considered efficient from a system perspective. This way, inefficient development of the system will at least be recognized.

Congestion management

A main problem with fixed system access tariffs is that may give rise to a situation in which market transactions lead to a load flow pattern that the network cannot physically accommodate. Therefore fixed system access tariffs need to be supplemented with a congestion management method that allocates the scarce network capacity. Congestion management methods, the subject of Chapter 10, may also provide long-term economic incentives to the generation market and to network operators. Unfortunately, the research showed that none of the available methods provides efficient incentives to both sides. Given the choice, congestion pricing methods (essentially variations of auctions) are to be preferred to remedial methods (such as redispatching) because they provide better incentives to generators. Because the networks are necessarily regulated as natural monopolies, it is easier to guide their long-term development through other means.

Policy advice

Considering:

- the risk of underinvestment and manipulation of price spikes in an energy-only market,
- the resulting social costs, and
- the additional benefits of capacity mechanisms in the form of stabilization of prices and a reduction of market power,

electricity markets should have a capacity mechanism.

Ideally, a capacity mechanism is implemented jointly by closely interconnected electricity systems. Where this is not possible, importing systems may implement one unilaterally. However, this requires a more complex solution, especially in decentralized markets, and would affect international trade. The best options in this case are different versions of reliability contracts.

With respect to the coordination of generation with the network, the most important policy recommendation is for a paradigm shift: instead of maximizing competition, the focus should be upon overall economic efficiency in which the benefits of competition and other economic incentives are weighed against economies of coordination. The

merits of a version of locational marginal pricing or zonal pricing should be considered for implementation in the European grid.

Samenvatting

Het uitgangspunt voor de liberalisering van de Europese elektriciteitsmarkten is dat het netwerk, als natuurlijk monopolie, gereguleerd dient te blijven worden, terwijl de productie van elektriciteit verzorgd kan worden door een concurrerende markt die geen specifieke regulering behoeft. Deze studie trekt deze laatste aanname in twijfel. De specifieke karakteristieken van het elektriciteitssysteem, zoals de moeilijkheid om elektriciteit op te slaan, de lange levenscyclus van productie-eenheden en de nog langere levenscyclus van elektriciteitsnetten, en de nauwe relaties tussen productie en de netwerken, zorgen ervoor dat de dynamiek van de elektriciteitssector afwijkt van die van andere sectoren. Gezien de hoge maatschappelijke en economische kosten van verstoringen van de elektriciteitsvoorziening dienen elektriciteitsmarkten zorgvuldig ontworpen te worden.

De centrale onderzoeksvraag van deze studie is:

Biedt het huidige ontwerp van de Europese groothandelsmarkten voor elektriciteit voldoende lange-termijnprikkels voor voorzieningszekerheid en economische efficiëntie, en zo niet, wat zijn de beleidsopties voor interventie?

Het onderzoek richt zich op twee aspecten van investeringen in productievermogen in Europese elektriciteitsmarkten: de kwantitatieve vraag of de markt in voldoende productiecapaciteit zal blijven voorzien en de kwalitatieve vraag hoe investeringen in productiecapaciteit gecoördineerd kunnen worden met de fysieke mogelijkheden van het elektriciteitsnet. Hoofdstuk 2 introduceert de onderzoeksvraag, bakent het terrein af en beschrijft de methode. Hoofdstuk 3 beschrijft het conceptuele model dat aan de analyse ten grondslag ligt.

Voorzieningszekerheid

De analyse van voorzieningszekerheid begint met een casusbeschrijving van de elektriciteitscrisis in Californië in 2000 en 2001 (hoofdstuk 4). Deze casus is een referentiepunt voor zowel het openbaar beleid, vanwege de grote en breed gedocumenteerde invloed die de crisis had, als voor wetenschappelijk onderzoek, vanwege het complexe maar fascinerende materiaal dat de casus oplevert. De conclusie is dat, ondanks de vele eigenaardigheden van het marktontwerp in Californië en de vele fouten die er zijn gemaakt, de onderliggende factoren die de crisis veroorzaakt hebben

niet uniek zijn voor Californië:

- Jarenlang bleven investeringen in productiecapaciteit achter bij de groei van de vraag, ook al bleken de productiebedrijven grote winsten te kunnen maken toen er zich tekorten ontwikkelden.
- De lage prijselasticiteit van de vraag, gecombineerd met de afwezigheid van langetermijncontracten, stelde producenten in staat om de groothandelsprijzen ver boven het niveau van een concurrerende markt te drijven door productiecapaciteit achter te houden. Dit droeg in aanzienlijke mate bij aan het optreden van stroomonderbrekingen en aan de hoge prijzen die de consumenten betaalden.
- De crisis was in gang gezet door een afname van importen uit naburige staten, die een verschillende marktstructuur hadden. Geconfronteerd met schaarste gaven deze staten hun eigen consumenten voorrang.

Hoofdstuk 5 analyseert waarom in concurrerende *energy-only* markten, waarin de elektriciteitsprijs de enige drijfveer voor investeringen is, niet altijd een sociaal optimaal volume aan productiecapaciteit tot stand komt. De lage prijselasticiteit van de vraag en de moeilijkheid om stroom op te slaan zijn er debet aan dat elektriciteitsprijzen in de meeste markten zeer volatiel zijn. Hierdoor zijn investeringen in productiecapaciteit risicovol. In theorie zouden risiconeutrale investeerders een sociaal optimaal volume aan productiecapaciteit creëren, maar beleidsonzekerheid en een gebrek aan transparantie in de markt kunnen het investeringsevenwicht gemakkelijk lager doen worden dan het sociale optimum.

Het grootste risico voor consumenten is niet zozeer statische onderinvestering als het risico op investeringscycli. De lange aanlooptijd voor nieuwe productiecapaciteit betekent dat de reactie van investeerders op prijsprikkels een aantal jaren vertraagd wordt. Zelfs als er gemiddeld voldoende geïnvesteerd zou worden, zouden investeringscycli hoge maatschappelijke kosten met zich meebrengen, zoals de elektriciteitscrisis in Californië gedemonstreerd heeft.

Als het optimale niveau van productiecapaciteit, dat natuurlijk verandert met de tijd, niet precies bereikt kan worden, is het belang van consumenten om eerder teveel dan te weinig te investeren. De kosten van overcapaciteit zijn beperkt in vergelijking tot de maatschappelijke kosten van tekorten. Bovendien worden de kosten van een beperkte mate van overcapaciteit tenminste gedeeltelijk gecompenseerd door een vermindering van de marktmacht van producenten. Om deze redenen hebben elektriciteitsmarkten een 'capaciteitsmechanisme' nodig, een systeem van regels en prikkels met als doel het volume aan productiecapaciteit te stabiliseren. Op dit moment hebben de meeste Europese elektriciteitsmarkten geen capaciteitsmechanisme.

Hoofdstuk 6 beschrijft een aantal bestaande en voorgestelde capaciteitsmechanismen. Hoofdstuk 7 ontwikkelt een analytisch kader voor het evalueren van capaciteitsmechanismen en past het toe op de capaciteitsmechanismen uit hoofdstuk 6. De eerste conclusie is dat capaciteitsmechanismen die direct ingrijpen op het volume aan productiecapaciteit (capaciteitsvereisten, betrouwbaarheidscontracten en capaciteitsabonnementen) de voorkeur verdienen over mechanismen die indirect werken door het gebruik van prijsprikkels.

Gegeven het internationale karakter van Europese elektriciteitsmarkten zou de beste oplossing zijn om een capaciteitsmechanisme gezamenlijk te implementeren. Dit is niet alleen van belang om verstoring van de internationale handel te voorkomen, maar ook omdat het lastiger is om een effectief capaciteitsmechanisme te ontwikkelen in een gedecentraliseerd elektriciteitssysteem als de naburige systemen niet een vergelijkbaar capaciteitsmechanisme hebben. Het is echter de vraag of het mogelijk is om dit doel te bereiken voordat de eerste schaarsteperiode van een investeringscyclus zich voordoet.

Individuele Europese landen kunnen ervoor kiezen om unilateraal een capaciteitsmechanisme te implementeren terwijl zij wachten op de ontwikkeling van een regionale oplossing. Helaas zijn geen van de capaciteitsmechanismen van hoofdstuk 6 effectief in een open, gedecentraliseerd systeem. Tijdens een regionaal tekort kan de productiecapaciteit die tenminste gedeeltelijk met het capaciteitsmechanisme gefinancierd is gebruikt worden om stroom te exporteren, waardoor de consumenten in het systeem met het capaciteitsmechanisme geen lagere prijzen of betere voorzieningszekerheid hebben dan zonder een capaciteitsmechanisme.

Hoofdstuk 8 onderzoekt enkele innovatieve oplossingen voor het specifieke, typisch Europese, geval van een open, gedecentraliseerd systeem. Twee varianten van betrouwbaarheidscontracten zijn de meest aantrekkelijke oplossingen voor de middellange termijn. De centrale variant heeft als nadelen dat hij complex is en gebruik maakt van veilingen die gevoelig zouden kunnen zijn voor het uitoefenen van marktmacht. De bilaterale variant heeft als nadelen dat hij minder effectief zou kunnen zijn bij verticale integratie van producenten met leveranciers. Op de lange termijn, als alle consumenten meters hebben die het tijdstip van het verbruik meten, lijkt een financiële versie van capaciteitsabonnementen een aantrekkelijk alternatief.

Coördinatie

Het tweede aspect van investeringen in productiecapaciteit dat onderzocht is, is de kwestie hoe de elektriciteitsmarkt te coördineren met het netwerk (hoofdstukken 9 en 10). Om een gelijk speelveld te scheppen voor de markt voor elektriciteitsproductie is het essentieel dat productiebedrijven 'ontvlochten' worden van de netwerkbedrijven, zodat de netwerkbedrijven geen economische belangen hebben in de productiemarkt. Fysiek bestaat er echter de noodzaak om de bedrijfsvoering en de ontwikkeling van productie en de netten te coördineren met betrekking tot:

- de bedrijfsvoering van productie-eenheden,
- de blindvermogenshuishouding, en
- de locatiekeuze van nieuwe productie-eenheden.

In een geliberaliseerde markt heeft de systeembeheerder geen directe autoriteit over operationele en investeringsbeslissingen met betrekking tot productiecapaciteit. Daarom was het ideaal dat de productiebedrijven geleid zouden worden door efficiënte economische prikkels. Dit betekent dat netwerktarieven niet alleen het dekken van de netwerkkosten als doel zouden moeten hebben, maar ook producenten prikkels zouden moeten geven om het netwerk op efficiënte wijze te gebruiken. Een mogelijke oplossing is het systeem van *locational marginal pricing* zoals dat op sommige plaatsen in de VS en in Nieuw Zeeland gebruikt wordt.

Europa heeft een andere marktstructuur gekozen, in ieder geval gedeeltelijk uit noodzaak: de institutionele vereisten voor *locational marginal pricing* lijken te hoog te zijn voor Europese landen, met hun uiteenlopende juridische, institutionele en economische structuren, om op korte termijn te realiseren. Bovendien wilde men de markt stimuleren door middel van een simpel en transparant systeem van netwerktoegang. Dit onderzoeksproject heeft de gevolgen onderzocht van de Europese keuze voor vaste netwerktarieven. De aanwezigheid van voortdurend veranderende netwerkexternaliteiten betekent dat de vaste netwerktarieven geen economisch efficiënte prikkels kunnen geven. Hierdoor belemmeren zij de coördinatie tussen de ontwikkeling van de netwerken en het productievermogen op de lange termijn.

Beleidsopties voor het verbeteren van operationele en lange-termijnprikkels voor productiebedrijven in systemen met vaste transmissietarieven omvatten het gebruik van congestiemanagementmethoden, variaties in de aansluittarieven, betalingen voor systeemdiensten en vergunningen. Een combinatie van dergelijke prikkels zal resulteren in een pragmatische mengelmoes, niet in een theoretisch consistent systeem. De mengelmoes van prikkels zal van tijd tot tijd opnieuw gekalibreerd moeten worden om het gewenste effect te bereiken. Hierdoor kan de paradoxale situatie ontstaan van een complexe, intransparante combinatie van *ad hoc* maatregelen om de negatieve bijeffecten van vaste netwerktarieven, die eenvoud en duidelijkheid als doel hadden, te compenseren.

Gezien het gebrek aan gegevens over dit onderwerp dient de ontwikkeling van de markten gemonitord worden met betrekking tot de efficiëntie van beslissingen door productiebedrijven aangaande investeringen in en het uit bedrijf nemen van centrales. Deze beslissingen dienen vergeleken te worden met wat efficiënt zou zijn voor het systeem als geheel. Zo wordt een inefficiënte ontwikkeling van het systeem in ieder geval gesignaleerd.

Congestiemanagement

Een belangrijk nadeel van vaste transmissietarieven is dat zij tot een situatie kunnen leiden waarin de transacties van de marktpartijen een belastingspatroon van het netwerk ten gevolge hebben die fysiek niet uitvoerbaar is. Daarom dienen vaste transmissietarieven aangevuld te worden met een systeem voor het managen van congestie waarmee schaarse netwerkcapaciteit toegewezen kan worden. Congestiemanagementmethoden, het onderwerp van hoofdstuk 10, kunnen ook lange-termijnprikkels geven aan de productiemarkt en aan netwerkbeheerders. Helaas wijst dit onderzoek uit dat geen van de beschikbare methoden aan beide groepen tegelijkertijd de juiste prikkels kan geven. Gegeven de keuze wordt het beprijzen van congestie (diverse varianten van veilingen) verkozen boven correctieve methoden (zoals *redispatching*) omdat ze betere prikkels geven aan producenten. Omdat de netwerken, als natuurlijke monopolies, noodzakelijkerwijze gereguleerd zijn, zijn er meer aangrijpingspunten om hun lange-termijnontwikkeling te beïnvloeden.

Beleidsadvies

Gezien:

- het risico op onderinvestering en manipulatie van prijspieken in een *energy-only* markt,
- de maatschappelijke kosten die daar het gevolg van zijn,
- de bijkomende voordelen van capaciteitsmechanismen in de vorm van stabilisering van prijzen en vermindering van marktmacht,

dienen elektriciteitsmarkten een capaciteitsmechanisme te hebben.

Idealiter wordt een capaciteitsmechanisme gezamenlijk ingevoerd door een zo groot mogelijke groep van met elkaar verbonden elektriciteitssystemen. Indien dit niet mogelijk is kunnen importerende systemen eenzijdig een capaciteitsmechanisme invoeren. Dit vergt echter een complexere oplossing, met name in gedecentraliseerde markten, en zou internationale handel beïnvloeden. De beste opties in dit geval zijn de verschillende versies van betrouwbaarheidscontracten.

Wat betreft de coördinatie van productie en het netwerk is de belangrijkste aanbeveling om het paradigma te veranderen: in plaats van het maximaliseren van concurrentie dient de nadruk te liggen op de economische efficiëntie van het systeem als geheel, waarbij de voordelen van concurrentie en andere economische prikkels afgewogen worden tegen de voordelen van coördinatie. De voor- en nadelen van *locational marginal pricing* of van een systeem van prijszones zouden onderzocht moeten worden met betrekking tot implementatie in het Europese netwerk.

Curriculum Vitae

Laurens James de Vries was born in Amsterdam on May 28, 1967. From 1979 to 1985 he attended the Maurick College high school (VWO). In 1985 he moved to Delft, where he obtained his Master of Mechanical Engineering Degree in 1991, with a specialization in environmental and energy technology. After his graduation he moved to Olympia, Washington State, USA, where he enrolled part-time in the Master of Environmental Studies program. Concurrently he held traineeships at several local government agencies, among others in the fields of energy conservation and drinking water supply, and in the last few years there he worked as an assistant planner in the Public Transportation and Rail Division of the Washington State Department of Transportation. He obtained his Master of Environmental Studies Degree from The Evergreen State College in 1996. In the summer of 1997 he married Deborah Sherwood. Their honeymoon led them through a large portion of Asia and lasted until the summer of 1998, when they settled in the Netherlands. The following year Laurens worked at NovioConsult in Nijmegen, an environmental consulting firm which specializes in advising local governments, after which he decided to return to academics. In 1999 he commenced his Ph.D. studies with the Department of Technology, Policy and Management of Delft University of Technology.