PEAKING CAPACITY IN RESTRUCTURED POWER SYSTEMS

Gerard L. Doorman

Norwegian University of Science and Technology
Faculty of Electrical Engineering and Telecommunications
Department of Electrical Power Engineering

A thesis submitted in partial fulfilment of
the requirements for the degree of
doktor ingeniør

URN:NBN:no-2136

PREFACE

All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident - Arthur Schopenhauer

The work with this thesis started in 1996 under the supervision of the late Professor Øivin Skarstein, whose untimely death shocked us all in the summer of 2000. Because of the direction this research took as time went by, supervision was taken over by Professor Ivar Wangensteen in the spring of 1999. Professor Bjørn Nygreen has been supporting supervisor in the field of operations research. All three have given their invaluable contributions to the final result of the work.

The research has been primarily financed by the "Effekt" program, funded by the Research Council of Norway and a number of organizations in the Norwegian electricity industry. My employer, SINTEF Energy Research, has contributed with additional funding.

The subject at hand is complicated and involves many angles of incidence from various disciplines: technical, economic, political and sociological, among others. The approach taken here has been from the technical and economic sides, but within neither of these fields, it has been possible to give all the answers. My hope is that the present work will contribute to a better understanding of the problem that has been studied, and that it will be a basis for the development of good practical solutions.

Apart from the three professors already mentioned, many people have contributed to this work. Taking the risk of forgetting some, I will at least mention:

- Sverre Aam, managing director and Nils Flatabø, research director, both of SINTEF Energy Research
- Many of my colleagues, and especially Anders Gjelsvik, Olav Fosso, Ståle Huse and Arne Johannesen
- Professor Michel Roland from the University of Laval, Canada, for significant contributions to Chapter 6
- My brother Maarten for some essential support during a difficult phase
- Last but not least, my wife Anbjørg and my sons Jan Henrik and Thomas for love and support

This work is dedicated to the memory of my father, who died suddenly a few months before this thesis was finished.

Trondheim, November 2000.

SUMMARY

Background

The theme of this thesis is the supply of capacity during peak demand in restructured power systems. There are a number of reasons why there is uncertainty about whether an energy-only electricity market (where generators are only paid for the energy produced) is able to ensure uninterrupted supply during peak load conditions.

Much of the public debate in Europe has been about the present surplus generation capacity. However, in a truly competitive environment, it is hard to believe that seldom used capacity will be kept operational. This is illustrated by developments in Sweden. For this reason, the large surplus of generation capacity in the European Union may vanish much faster than generally assumed. In the USA, much of the debate has been about California. During the last three summers, California has occasionally experienced involuntary load shedding and prices have been very high during these periods. To some extent, the Californian situation illustrates the relevance of the subject of this thesis: in a deregulated system generators may not be willing to invest in peaking capacity that is only needed occasionally, even though prices are very high during these periods.

A good solution to the problem of providing peaking power is pivotal to the success of power market restructuring. Solutions that fail to create the right incentives will result in unacceptable load shedding and can endanger the whole restructuring process. On the other hand, solutions that pay too much for investments in peaking power will lead to generation capacity surpluses and thus represent a societal loss.

Why is peaking capacity a problematic issue in energy-only markets?

Traditionally, probabilistic methods are applied to calculate the required generation capacity to obtain a desired level of reliability. In a centrally planned system, this level of generation capacity is developed in a least-cost manner. A single utility or central authorities can thus control the level of reliability directly. This is not possible in a market-based system, if suppliers are only paid for the energy produced.

Under the assumption of certainty and continually varying prices, generators fully recover their variable and investment costs under ideal market conditions. When uncertainty is taken into account, generators will cover their expected costs. However, revenues will be extremely volatile, especially for peaking generators. Combined with a risk-averse attitude, it is unlikely that investments will be sufficient to maintain the traditional level of reliability in an energy-only market. Consequently, one would expect reserve margins to decline in such markets. This effect is very clear in Sweden that deregulated in 1996, and less explicit in a number of other cases like Norway, California and Alberta.

Pricing and Consumer Preferences

The theory of electricity pricing was originally developed for vertically integrated utilities, but elements from this theory are also valuable in a restructured context. Many authors have agreed on the presence of a capacity element in the optimal price during peak-load conditions, while price should equal marginal cost during low-load conditions. An important assumption is that prices have to be stable. More recently, spot pricing of electricity has been advocated. A number of papers have been written about how to efficiently include security considerations in the spot price.

Because the availability of capacity cannot be directly controlled in an energy-only spot market, the probability of occasional capacity shortages increases. It is important to be prepared for this situation. The core of the problem is that demand is de facto inelastic in the short-term because of traditional tariff systems. It is shown that considerable economic gains are obtained when demand elasticity can be utilized, even if only minor shares of demand are elastic in the short-term. Better utilization of demand elasticity was also profitable in traditional systems, but after restructuring the gain is much larger: the alternative is not expensive generation but random rationing, which is unacceptable in modern society.

It is possible to go one step further. Consumers have different preferences for the use of energy and reliability. Some consumers have a low tolerance about being disconnected, while others are more willing to accept this. This will be reflected by their willingness to pay for reliability. A better solution would emerge if consumers could buy electricity and reliability more or less as separate commodities, based on their preferences.

In the context of pricing it should be pointed out that “profile-based settlement” that allows small consumers to freely choose their supplier without hourly metering is detrimental with respect to the correct pricing of capacity. It should only be used in the initial phases of opening a market.

Improved utilization of system resources

Even in the short-term, demand and the availability of generation and transmission resources are uncertain. Therefore, it is necessary to have reserves available in a power system. When capacity becomes scarce, it is difficult to satisfy the reserve requirements. If these requirements are strict, the only possibility is to resort to what can be called “preventive load-shedding” to satisfy the reserve requirements. This is obviously an expensive solution, but there are no obvious ways of balancing the (societal) cost of preventive load shedding against reduced system security. In this thesis, a model is developed for unit commitment and dispatch with a one-hour time horizon, with the objective of minimizing the sum of the operation and disruption costs, including the expected cost of system collapse. The model is run for the IEEE Reliability Test System. It is shown that under conditions where there is not enough capacity available to satisfy the reserve requirements, large cost savings can be obtained by optimizing the sum of the operation and disruption costs instead of using preventive load-shedding. In the model, it is also possible to directly target reliability indexes like the Loss of Load Probability or Expected Energy not Served. It is shown that increased reliability (in terms of the values of the indexes) can be obtained at a lower cost by targeting the indexes directly instead of resorting to reserve requirements. This is especially the case if flexible load-shedding routines are developed, making it possible to disconnect and reconnect the optimal amounts of load efficiently.

The use of alternatives to fixed reserve requirements as a means to maintain system security does not solve the problem about how to ensure the availability of peaking capacity. However, in a situation with occasional capacity shortages, it gives the System Operator a tool to find the optimal balance between preventive load shedding and system security, which can result in significantly lower disruption costs in such cases. More research and development in this area is necessary to develop methods and tools that are suitable for large power systems.

Ancillary Services

Investment in peaking capacity is insufficient in restructured systems because expected revenues are too low or too uncertain. If generator revenues are increased, the situation improves. One way to obtain this is to create markets for ancillary services. In the thesis, a model is developed for a central-dispatch type of pool. In this model, markets for energy and three types of ancillary services are cleared simultaneously for 24 hours ahead. Market prices are such that volumes and prices are consistent with the market participants' self-dispatch decisions – i.e. given these prices, market participants would have chosen the same production of energy and ancillary services as the outcome of the optimization program. With this model, it is shown that markets for ancillary services increase generator revenues, but this effect is partly offset by lower energy prices. This shows that markets for ancillary services can contribute to improving the situation, but given the remaining uncertainty, this is hardly enough to solve the problem.

Capacity Subscription

Because consumers have preferences for two goods: electricity and reliability, they should ideally have the choice of purchasing the preferred amount of each of these. Traditionally this is not possible – reliability is a public good, produced or obtained by a central authority on behalf of all consumers. Technological progress is presently changing this. Capacity subscription is a method that allows consumers to choose their individual level of reliability, at the same time creating a true market for capacity. It is based on the concept of self-rationing. Consumers anticipate (for example on a seasonal basis) their need for capacity *at the instant of system-wide peak demand*. Based on this anticipation, they procure their desired level of capacity in a market, where generators offer their available capacity. Demand is limited to subscribed capacity by a fuse-like device that is *activated when total demand exceeds total available generation*. In this way, the capacity payment only influences the market when demand is close to installed capacity, and does not distort the energy price in other periods. Demand is not limited when there is ample capacity. Demand will never exceed supply, because it can be limited in an acceptable way when this situation occurs. Moreover, both consumers and suppliers can adapt to situations with scarce or ample capacity, and the price of capacity will reflect this situation. There is one problem with the method: as consumers do not reach their subscribed capacity simultaneously, there will be a capacity surplus at the instant the fuse-devices are activated. Two methods to solve this problem are analysed, and it is shown that the problem can be solved optimally by giving consumers who prefer this the opportunity to buy power in excess of their subscription on the spot market.

Policy evaluation

Six alternative policies to assess the peaking power problem are analysed based on the following criteria:

- Static efficiency: the welfare-optimal match of consumption and supply
- Dynamic efficiency: the ability to create incentives for innovation
- Invisibility: with invisible strategies, each market actor pursues his or her own objectives without worrying about anyone else's
- Robustness: a robust policy is less sensitive to deviations from assumptions
- Timeliness: the ability of a policy to be employed at the right time
- Stakeholder equity: the degree to which all the involved parties are treated equitable
- Corrigibility: the extent to which a policy can be corrected once it is employed
- Acceptability: the degree to which the policy is acceptable to all parties
- Simplicity: *ceteris paribus* simple strategies are preferable over more complicated strategies
- Cost: the cost of implementing the policy
- System security: the policy's ability to obtain an acceptable level of system security

The policies are, in short (an example is given in parentheses):

- Capacity obligation: suppliers are obliged to keep sufficient capacity (PJM)
- Fixed capacity payment: a fixed payment is offered for available capacity (Spain)
- Dynamic capacity payment: capacity payment is based on the Loss of Load Probability (England and Wales)
- Energy-only: no explicit payments or obligation (Scandinavia, California)
- Proxy prices: very high administrative prices are used as a proxy to the Value of Lost Load when load shedding is necessary (Australia)
- Capacity subscription: cf. the description above (not implemented)

As could be expected, no single policy performs best on all criteria. The obligation and fixed payment methods do not perform well on market efficiency criteria, as essentially they are not market-based policies. The proxy prices policy is a reasonable policy on most criteria. It is easy, cheap and quick to implement. Because there is little experience with the method so far, there is some uncertainty with respect to if it is effective. One can anticipate that the threat of having to buy power at rationing prices will motivate market participants to avoid coming in a buying position in such cases, and that this will stimulate the adaptation of innovative solutions, especially on the demand side.

The capacity subscription policy looks very promising on the issues of efficiency, robustness and system security. This is especially true for dynamic efficiency: consumers will weigh the cost of capacity against the cost of innovative load control devices, and if the price of capacity is high, a market for such technology will emerge. However, there is a considerable threshold prior to the introduction of capacity subscription, caused by the implementation costs and complexity.

The conclusion on policies is thus that in an early stage after restructuring it may be appropriate to resort to the capacity obligation or payment method if the capacity balance is tight at the time of transition. For the medium-term, or if there is ample capacity initially, it is sensible to introduce proxy market prices to transfer the risk of a capacity deficit to market participants, with due attention being paid to the appropriate price level. Capacity subscription can be a long-term objective.

TERMS AND ABBREVIATIONS

AGC	-	Automatic Generator Control
AMPL	-	A Mathematical Programming Language
B&B	-	Branch and Bound
BTU	-	British Thermal Unit
CC	-	Capacity Charge
CCGT	-	Combined Cycle Gas Turbine
CEGB	-	Central Electricity Generation Board (the pre-restructuring utility in the UK)
CHP	-	Combined Heat and Power
CPLEX	-	Solver for mathematical problems
DP	-	Dynamic Programming
EENS	-	Expected Energy Not Served
EIR	-	Expected Index of Reliability
F&D	-	Frequency and Duration
FERC	-	Federal Energy Regulatory Commission
FOR	-	Forced Outage Rate
GWh	-	Gigawatthour (=1000 MWh)
HLI	-	Hierarchical Level I (generation)
HLII	-	Hierarchical Level II (generation & transmission)
HLIII	-	Hierarchical Level III (generation & transmission & distribution)
IEA	-	International Energy Agency
IEEE	-	The Institution of Electrical and Electronic Engineers
ISO	-	Independent System Operator (does not own the transmission system)
IP	-	Integer Programming
IPP	-	Independent Power Producer
KKT	-	Karoush-Kuhn-Tucker
KvF	-	Kraftverksföreningen (the Swedish Power Association)
LOLE	-	Loss Of Load Expectation (days/year)
LOLP	-	Loss Of Load Probability
LR	-	Langrangian Relaxation
LRMC	-	Long Run Marginal Cost
MBTU	-	Million BTU
MC	-	Marginal Cost
MTTF	-	Mean Time To Failure
MWh	-	Megawatthour (= 1000 kWh)
MO	-	Market Operator
NGC	-	National Grid Company
NEPOOL	-	New England Power Pool
NERC	-	North-American Reliability Council
NOK	-	Norwegian Krone

Nordel	-	The organization for Nordic power cooperation
NordPool	-	The Nordic Market Operator
NPV	-	Net Present Value
NVE	-	“Norges Vassdrags og Energidirektorat”, (the Norwegian Directorate for water resources and electricity)
ORR	-	Outage Replacement Rate
PJM	-	Pennsylvania, Jersey and Maryland
PPP	-	Pool Purchase Price
PV	-	Present Value
PX	-	Power Exchange
RTS	-	Reliability Test System
SEK	-	Swedish Krona
SEP	-	Samenwerkende Electriciteits Productiebedrijven (the pre-restructuring body for cooperation between Dutch generators)
SMP	-	System Marginal Price
Statnett	-	The Norwegian grid owner and System Operator
SRMC	-	Short Run Marginal Cost
Svenska Kraftnät	-	The Swedish grid owner and System Operator
TSO	-	Transmission System Operator
TWh	-	Terawatthour (= 1000 GWh)
UC	-	Unit Commitment
UCTE	-	Union for the Co-ordination of Transmission of Electricity (in Europe), formerly UCPTE
VOLL	-	Value Of Lost Load

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

The theme of this thesis is the supply of capacity during peak demand in deregulated or, rather, restructured[1] power systems. It may seem odd that this should be a problem. After all, the point of restructuring is to substitute central planning by a market-based structure. In the new environment, it should not be necessary for central authorities or planning bodies to bother about supply: the market should ensure an efficient balance between supply and demand at any point in time.

However, there are a number of reasons why there is well-founded doubt about the ability of an energy-only electricity market (where generators are paid only for the energy produced[2]) to ensure uninterrupted supply during peak load conditions. This is partly the theme of this introductory chapter, and the main theme of the next chapter.

Section 1.1 places the theme of this thesis in the context of the present debate on power system restructuring. As a general background, Section 1.2 discusses the motivation for power system restructuring. This is a very broad item in itself, and the short description here will only serve as a reference for the remainder of the thesis. In Section 1.3, the central elements of power system restructuring will be outlined. Section 1.4 shows that a number of authors have been engaged in the question of the provision of peaking power after restructuring. In that section, the problem is also discussed in some more detail. Section 1.5 shortly outlines the structure of the remainder of the thesis.

1.1 Peaking capacity and th e present debate on power system restructuring

Apart from some countries that started early (e.g. the UK, Scandinavia), the restructuring process in Europe is strongly influenced by EU Directive 96/92/EC on common rules for the internal market in electricity, which came into force in February 1997. The Directive gives the minimum requirements the EU member states have to satisfy with relation to power market restructuring. Some countries aim at satisfying the minimum requirements (e.g. France), while others go much further in the direction of liberalization (Spain, the Netherlands). Much

[1] A note on terminology is appropriate. The terms “deregulation”, “liberalization”, “restructuring”, and even “privatization” are used somewhat arbitrarily by different authors for either similar or different concepts. The proper term for the ongoing process, is “restructuring”, or possibly the more artificial “reregulation”. As pointed out by Jaffe and Felder [1.6], “Government can and must establish the rules under which competition will occur. (…) The notion that we can “get government out of the way and let the market rip” is devoid of serious meaning”. This observation somewhat disqualifies “deregulation”. “Liberalization” also understates the need for a comprehensive regulatory framework. As to privatization, this will often but must not necessarily (cf. the Nordic experience) accompany restructuring.

[2] There may be payments for other (ancillary) services as well in an energy-only market. The essential characteristic is that payments are basically short-term, and that there are no additional instruments to remunerate available capacity.

of the public debate has been about the present surplus generation capacity in Europe. In this situation, there is no imminent threat of a capacity deficiency, and the subject of this thesis may not seem very relevant. However, in a truly competitive environment, it is hard to believe that existing capacity that is hardly ever used is kept operational in the longer run. The situation in Sweden illustrates this, as will be shown later. For this reason, the large surplus of generation capacity in the European Union may vanish much quicker than generally assumed[3]. If that happens, the theme of this thesis will become very relevant.

In the United States, the drive for the restructuring process comes from FERC Order 888 of June 1996 ("Promoting Wholesale Competition Through Open Access; Non-discriminatory Transmission Services by Public Utilities; Recovery of Stranded Costs by Public Utilities and Transmitting Utilities"). Like in Europe, the speed of the process is very different in the various states. It has come furthest, and has created most debate in California, where the new electricity market was launched on 1 March 1998. The Californian market structure is complicated, and has been adjusted more or less continually since its introduction. During the last three summers, California has experienced several periods of very hot weather and consequently high demand. During these periods it has occasionally been problematic to provide sufficient reserves, and involuntary load shedding has occurred. Moreover, prices have been very high during these periods. This situation has created considerable debate about the viability or at least desirability of the reform. To some extent[4], the Californian situation illustrates the relevance of the subject of the thesis: in a deregulated system generators may not be willing to invest in peaking capacity that is needed only occasionally, even though prices are very high during these periods.

This thesis will provide a number of arguments why the simplest market organization, where generators are paid only for the energy produced, does not create the incentives for investment in peaking power. Following this argumentation, it will analyse a number of ways to improve this situation. Several market organizations that have additional instruments will be compared.

A good solution to the problem of providing peaking power is one of the keys to the success of power market restructuring. Solutions that fail to create the right incentives will result in unacceptable load shedding and claims to abandon the whole restructuring process.

[3] The closing of unprofitable plants can be seen as an indicator of the degree of real competition in this market. As long as these plants are kept operational, without creating any profits for their owners, the motivation to cut costs is clearly not too high, indicating a lack of real competition.

[4] Only to some extent, because in California there is clearly not one single cause of the problems. In contrast to Europe, demand is still rising considerably in California, but partly due to tough environmental laws, no new plants have been built in the last decade [1.1]. New entrants are burdened with part of the cost of "stranded assets" built by incumbents, making it difficult for them to compete effectively on price. Finally, price caps are used frequently, creating additional uncertainty for investors in new capacity.

On the other hand, solutions that overcompensate investments in peaking power will lead to generation capacity surpluses, representing a societal loss. The impact of the latter situation may seem less dramatic than that of the former, but unnecessary investments of hundreds of millions, possibly billions of USD can be avoided by implementing the right policy.

The objective of this thesis is on the one hand to exploit ways to reduce the impact of occasional capacity shortages and on the other hand to analyse alternative policies to reduce the probability of such situations occurring. The realm of the work is within the power sectors of highly industrialized countries with low growth of electricity demand. The situation is clearly different in high-growth countries, where the major concern is to obtain sufficient investment in new capacity, while the societal losses of (limited) overinvestment are small - they will be overtaken by next year's growth.

The problem of covering peak demand is not unique for the power market, but is seen in many other situations as well: road queues during rush hours, taxis on New Year's Eve, charters to attractive destinations on the first day of a holiday, telecommunication systems etc. However, a power system has some specific characteristics that require special solutions.

1.2 The motivation for power market restructuring

Starting in Chile in 1978 [1.2], power market restructuring has now become a world-wide phenomenon both in developed and developing countries. A good overview over the background for this development is given by Paul Joskow in e.g. [1.3], [1.4] or [1.5]. Joskow recognizes the following motives:

Generation construction costs and investment decisions. There is substantial variation across utilities in the construction costs of similar generating units, which cannot be explained by differences in underlying cost opportunities[5].

Operating cost of generating units. A traditionally organized[6] electricity sector offers a great opportunity for the political system to pursue other objectives, e.g. the use of domestic fuels. According to Newberry [1.7], "… competition is the key to improving the efficiency of the electricity industry, in part because it makes it more difficult for government or

[5] Joskow gives three references to sustain this case. Tenenbaum et al. give some examples of this in their note 3 in [1.11]. No research has been done on this topic in Norway to this author's knowledge, possibly because such comparisons are much harder to conduct for hydropower projects.

[6] The term "traditional organization" is used for the broad spectre of organizational structures dominating electricity sectors around the world before the restructuring process started. Often "vertical organization" is used, but this does not completely cover all the different forms of traditional organization. The Norwegian system, for example, has always consisted of a combination of vertically integrated utilities, a great number of distribution companies and a number of clear-cut power producers. This is far from the single-utility organization in France and in the United Kingdom before 1990.

bureaucrats to intervene". This view is also supported by Tenenbaum et al. "Subsidies are easier to sustain when concealed in the purchases or operating expenses of government owned electricity enterprises". Moreover, some utilities appear to be systematically better operators than others. Finally, utilities may have incentives to keep generating units operating even when they should be closed.

Employment practices and wages. Experience after restructuring suggests that public enterprises and private firms subject to price and entry regulation employ too many workers (have too low levels of labour productivity).

Pricing inefficiencies: In traditionally organized systems, prices are often poorly aligned with the relevant marginal costs, because tariffs are designed to recover costs. As a result, when there is excess capacity, prices tend to rise, and when capacity is short, prices tend to fall, the opposite of how a market would work.

Innovation. Experience from both the UK[7] and the US suggests that when the restructuring process starts and entry barriers are removed, innovation in generation technology is stimulated.

Related to the first point there is also the problem of "excessive costs and prices" [1.5] in developed countries, due to overinvestment[8,9]. The opposite problem, the inability of the electricity sectors to attract adequate investment capital, occurs in developing countries. It looks contradictory at first sight that opposite problems (over- and underinvestment) should be solved by the same remedy (restructuring). The common denominator, however, is an inefficient allocation of capital under traditional systems.

The motivation for the Norwegian restructuring shares many common properties with the general points given above [1.13]:

- A questionable order in the development of hydropower resources (more expensive projects before cheaper ones)
- Cross-subsidizing in vertically integrated utilities
- Regional obligations to supply had resulted in large price differences
- Discrimination of small consumers compared with large industrial users

[7] The restructuring of the power system in England and Wales is generally referred to as "the British deregulation" or "the deregulation in the UK". This may be due to poor geographical knowledge or limited acquaintance with the geographical spread of the reform. The most probable reason, however, is linguistic: England and Wales is evidently much more bothersome to write than the alternatives. In this thesis all three variants will be used.

[8] Cf. Newberry [1.7], "…, most privatizations in mature markets start with excess capacity and considerable slack", Joskow [1.4], "…, the US industry has been plagued by excess capacity for most of the last two decades" (in spite of the fact that the US always has had a high share of private ownership) or Unda regarding the Spanish system [1.8].

[9] The magnitude or, ultimately even the presence of excess investment is not straightforward to demonstrate, due to the necessity of keeping sufficient reserves, and a potential disagreement on what is "sufficient". This will be discussed in much more detail later in this thesis.

- Regulated import and export, impeding an efficient international power trade
- Mixed and unclear objectives for the power sector as a whole (the use of "energy policy" for other goals like regional settlement, employment, regional policy, industry policy etc)

By creating a new market structure where generators can compete at wholesale and retail free from regulation, there should be good chances that at least some of the inefficiencies associated with the institution of regulated monopoly will be reduced under the pressure of competition[10].

1.3 Central elements of restr ucturing

Although there is a huge variation between how power sectors are restructured around the world, a credible deregulation at least has to include the following elements:

- The generation sector is potentially competitive, and must be organized as separate business entities. To obtain an adequate level of competition, the number of generating companies should be large enough. How many is enough mainly depends on a number of external factors, which are outside the scope of this section. Depending on past practice in the country concerned, privatization will often accompany this process.
- The transmission network and probably some ancillary services[11] are natural monopolies[12] and provide a critical platform upon which competition among generators depends. The transmission network must be operated by an entity that is independent from the generation and distribution companies.
- Approaches to transmission pricing must be adopted that allow equal access to the transmission network for all parties.

In case the distribution companies continue to supply a bundled product (essentially energy and transport) to retail consumers that they serve exclusively in a geographic area, one refers to a "*wholesale competition model*" or "*portfolio model*".

[10] Of course, new inefficiencies may occur, due to shortcomings in the new structures. An example is the low level of competition in the UK market. This is documented, among others, by Green and Newberry in [1.9] and Newberry in [1.7]: "All the gains were reaped by shareholders, and the reason is that the price of electricity did not fall anything like as much as the cost of fuel or the reduction in non-fuel costs…". Another good argument for this viewpoint is given by Wolak et al.: "…., these two generators (National Power and PowerGen) are able to obtain prices for their output substantially in excess of their marginal costs of generation" [1.10].

[11] The definition of ancillary services may be given as "activities that pertain to the provision of all electric services necessary for efficient and reliable generation, transmission and delivery of active power with a sufficiently stable frequency and voltage".

[12] A natural monopoly is described in [1.14] as an industry where "Average cost is diminishing over a broad range of output levels, and minimum average cost can only be achieved by organizing the industry as a monopoly". A more thorough treatment may be found in [1.15].

If the "wire business" of the distribution companies also has a natural monopoly character, and is organized as a number of independent entities, retail customers buy limited "wires" services from the local distributor and arrange for their own power supplies directly with generators or competing suppliers. This case is referred as a "*retail competition model*".

During the last decade, extensive literature on restructuring issues has emerged. Important issues that have been discussed thorougly are:

Transmission pricing and access. As indicated above, a transparent, well-functioning transmission pricing system, giving equal access to all parties, and preferably with a proper incentive structure, is an absolute prerequisite to obtain real competition[13].

Transmission congestion management. This may be viewed as a special case of transmission pricing, but it is an extremely important issue in its own merit.

Short-term operation and ancillary services. Many authors seem to agree that the short-term operation of power systems in developed countries was reasonably efficient before deregulation[14]. The challenge is to retain efficient operation subject to a large array of constraints in a multi-owner environment.

Stranded assets. This issue addresses the question of if and/or how to compensate (the owners of) generation assets that become uneconomic in a restructured environment, due to their unfavourable cost structure. This has primarily been a considerable issue in the US, though remarkably less in Europe[15].

Regulatory issues. The question of how to ensure an effective regulation of the parts of the system that will continue to operate as monopolies.

Although many authors express a positive view of power system restructuring, some critical voices can also be heard. As an example Walker and Lough [1.16] conclude that the British, Norwegian, Argentinean and Chilean restructurings provide little support for a similar process in the US. This is partly based on their observation that US electricity prices are low already, partly because the US industry to a large degree has private ownership.

[13] The Norwegian Energy Act came into function from 1 January 1991. However, first when the issue of transmission tariffs was solved and introduced in May 1992, real competition emerged.

[14] For example from [1.4] "… the electric power networks in most developed countries operate with very high levels of reliability and, given the short-run operating costs and availability of generators connected to the system, come reasonably close to efficient generator dispatch."

[15] In Sweden, some thermal generating capacity has been closed down without much discussion, at least with respect to stranded assets. Very little discussion on the subject in the UK has emerged in international forums. The different focus in the US and Europe may be related to a different starting point: while generation capacity was almost exclusively in public hands in Europe, private investors own more than 70 % in the US. The stranded costs issue is much easier to solve as a part of the privatization process, than as part of a restructuring transition in an already privatized industry. This argument is supported by the remark of Tenenbaum et al. in [1.11] that "… it is infinitely easier to make changes before privatization than to do so afterwards"

Another critical view is expressed by Casazza in [1.17]. His main conclusion is that the changes in the UK have been harmful to consumers, but extremely beneficial to shareholders. Also Casazza's main points against an American implementation of the British model are that the industry in the US already is in private hands to a large extent, and that prices are low.

These are contestable views. Even if prices generally may be low, there are great regional variations in the US, which was one of the motives for the Norwegian process. In addition, Newberry [1.7] argues that competition, not privatization increased efficiency in the UK.

1.4 Why is peaking power an issue in restructured systems?

Despite the lack of focus so far, some authors and observers have questioned if "the market" will provide enough generation capacity in the long run. Some examples:

- In [1.11] Tenenbaum et al. presuppose that "Customers that withdraw from the franchise (their local distributor) must be required to procure sufficient generating capacity to ensure their security of supply. (…) In other words, a regime that permits customers to purchase energy without any capacity responsibility is unfair to those customers who purchase the capacity that provides emergency protection for everyone connected to the grid". These authors characterize this as a "network externality", implying that they do not believe the market will provide the actual good (capacity) without intervention.
- In [1.22], Wangensteen and Holtan propose to establish a market for reserve capacity, based on assumptions that the market will not provide sufficient reserves in the long run.
- Fink and Van Son [1.23] argue that "Barring strong incentives to do otherwise, they (independent power producers) could very well prefer high power factor units designed for base load use, in preference to more expensive low power factor units. (…) Base load use means very constrained response capability. It will almost certainly be necessary, in contracts establishing access by producers to the bulk power network, to require that generators be fitted with adequate, active, well maintained governors, otherwise this expense is unlikely to be accepted by them". Clearly, the authors do not believe sufficient reserve capacity will be available without special measures.
- In [1.24], the World Energy Council states: "It remains to be seen, therefore, whether a liberalised market will be able to support necessarily long-term investment (…)".
- In a comment in the January 1997 issue of *Modern Power Systems*, the editor cites that the US "Department of Energy believes that the real test of the new market will be delayed until new capacity is needed and how high prices will have to get to stimulate investment".
- In a publication of KvF (the Swedish Power Association), Bo Källstrand, general manager of Graninge, a medium-sized Swedish producer, expresses: "Power producers can no longer bear the extra costs for increased reliability that these units (thermal reserve units being phased out) represent, because no compensation for these costs is offered" [1.27].
- Several market solutions (notably the UK [1.10], Argentina, Chile [1.2], New England [1.25] and the PJM Interconnection [1.26]) have included capacity payment or obligation elements. This is no proof that this is a necessary measure, but indicates that the designers

of these markets have concerns that enough capacity would not be provided by the market without these elements.

In addition, a few recent incidents indicate that the issue deserves more attention:

- Four times during the winter 1998/99 Statnett, the Norwegian System Operator had to take special measures to assure a sufficient level of reserves in the Norwegian market. In a number of other occasions the system was operated with reduced secondary reserves level, albeit with a considerable export. These occasions occurred under temperatures that by no means could be characterized as extreme[16].
- On Thursday 28 January 1999, Svenska Kraftnätt, the Swedish System Operator decided to forbid exports to Germany to ensure a satisfactory reserve margin in the Swedish system.

These incidents are indications that generation capacity may not be sufficient under all circumstances in the Norwegian and Swedish systems. Traditionally, the problem has been to satisfy demand at the minimum cost at any point in time during the planning horizon. In a restructured system, market agents have to make an investment that becomes profitable only if the spot price becomes high enough for at least a necessary number of hours. Due to a number of reasons, many mature power systems have had excess capacity (cf. footnote 8). So far, this fact has hidden the special need for short-term peaking power.

In this context, it is also appropriate to shortly discuss the concept of “capacity deficiency”. In the traditional power engineering view, to be discussed in Section 2.2, there is a capacity deficiency when available capacity is insufficient to cover demand with an acceptable level of reliability. From the view of economic theory, the whole concept of capacity deficiency is curious: the market price should settle at a level that implies a balance between supply and demand. However, there are two reasons why this may not happen:

- the public good character of reliability
- the majority of consumers buy electricity at a price that is fixed in the short run

Both these reasons will be discussed in this thesis. For now, it can be stated that as a result of these conditions, a capacity deficiency may occur. In this context, it will be defined as follows:

> A **capacity deficiency** occurs when the power system is unable to satisfy demand with agreed standards of reliability, resulting in *involuntary* load shedding.

[16] The Norwegian system has a high penetration of electrical heating, resulting in demand peaks during cold periods.

The resulting involuntary load shedding is the essential part of the definition. In a restructured system, conditions will change considerably. It is very probable that a more dynamic interaction between supply and demand will develop, resulting in various ways to limit demand under peaking conditions. As long as the limiting of demand is price based and consequently voluntary, there is no capacity deficiency, although installed capacity may be considerably lower than in traditional systems[17]. But when existing agreements are insufficient to secure a market balance, involuntary random load shedding may be necessary to maintain system reliability, and there is a capacity deficiency.

1.5 The structure of the thesis

The theme of this thesis is broad, and must be analysed from different angles. The theoretical foundations are found in the fields of economics and power engineering, while operations research has been an invaluable tool in several of the analyses that will be described.

The objective of Chapter 2 is to make plausible that the provision of peaking capacity in restructured systems actually is a problem. This is discussed from the viewpoints of power system reliability and finance theory, and illustrated with experience from some currently restructured systems.

In Chapters 3, 4 and 5 the problem is analysed from three different viewpoints:

- In Chapter 3, the starting point is pricing theory in traditional systems, which appears to have considerable relevance. From this starting point, a natural move to consumers and their preferences is made. The objective of the chapter is to show the importance of a real inclusion of the demand side in the market to obtain efficient results.
- In Chapter 4, a completely different viewpoint is taken. Reduced availability of peaking capacity challenges the system operator's ability to operate the system with satisfactory reliability. In power system operation, it is necessary to maintain a certain level of reserves to cope with uncertainties in demand and the availability of generation. There can be little doubt about the necessity of reserves, but their level and the corresponding level of reliability is or at least should be a point of discussion, cf. [1.28]. In Chapter 4, a model is developed to analyse the relation between reserves and reliability, and explore the possibilities of reducing reserve levels, while maintaining acceptable levels of reliability. In addition, the effects of a more flexible demand side are explored, creating a link to Chapter 3.
- In Chapter 5, a third approach to the problem is taken. An important role of excess capacity is to provide reserves, which in restructured systems are accommodated by so-called ancillary services. One can hypothesize that market-based payment for ancillary services increases the revenue from installed capacity, and create better incentives to make investments in capacity. To explore this hypothesis, a model is developed to compute

[17] In fact, this may be one of the major benefits of restructuring.

market prices for energy and ancillary services in a perfectly competitive market, and a number of analyses are done with the model, including the use of flexible demand

Partly based on the analysis in Chapter 3, the innovative concept of capacity subscription is introduced in Chapter 6. This concept is based on the theory of self rationing, that is well known from the economic literature on rationing. It is shown that this concept can be economically efficient in restructured markets, and that it will solve several of the problems at hand.

In Chapter 7, a comparison of several concepts to solve the peaking capacity problem is made, while Chapter 8 gives the conclusions of the work.

1.6 References

[1.1] "A shocking backlash", *The Economist*, August 26th 2000

[1.2] Hugh Rudnick, Ruy Vareal, William Hogan, "Evaluation of alternatives for power system coordination and pooling in a competitive environment", *IEEE Transactions on Power Systems*, Vol. 12, No. 2, May 1997

[1.3] Paul L. Joskow, Richard Schmalensee, "Markets for power: an analysis of electric utility deregulation", Cambridge, Mass. MIT, 1983

[1.4] Paul L. Joskow, "Introducing Competition into Regulated Network Industries: from Hierarchies to Markets in Electricity", *Industrial and Corporate Change*, No.2, 1996, pp. 241-282

[1.5] Paul L. Joskow, "Electricity Sectors in Transition", *The Energy Journal*, Vol. 19, No.2, 1998, pp. 25-52

[1.6] Adam B. Jaffe, Frank A. Felder, "Should Electricity Markets Have a Capacity Requirement? If So, How Should It Be Priced?", *The Electricity Journal*, December 1996, pp. 52-60

[1.7] David M. Newberry, "Freer electricity markets in the UK: a progress report", Keynote address to the 21st IAEE Annual International Conference Experimenting with Freer Markets, Quebec City, 15 May 1998, *Energy Policy*, Vol 26, No 10, 1998, pp. 743-749

[1.8] Juan Ignacio Unda Urzaiz, "Liberalization of the Spanish Electricity Sector: And Advanced Model", *The Electricity Journal*, June 1998, pp. 29-37.

[1.9] Richard J. Green, David M. Newberry, "Competition in the British Electricity Spot Market", *Journal of Political Economy, Vol. 100, No. 5, 1992.*

[1.10] Frank A. Wolak, Robert H. Patrick, "The Impact of Market Rules and Market Structure on the Price Determination Process in the England and Wales Electricity Market", UCEI Working Paper, February 1997 [cited 2000-10-30]. Available from Internet <http://www.ucei.berkeley.edu/ucei/PDFDown.html>

[1.11] Bernard Tenenbaum, Reinier Lock, Jim Barker, "Electricity Privatization. Structural, competitive and regulatory options", *Energy Policy*, Vol. 20, No. 12, December 1992, pp. 1134-1160

[1.12] H. Averch, L. Johnson, “Behavior of the firm under regulatory constraint”, American Economic Review, Vol. 52, 1962, pp. 1052-69

[1.13] Torstein Bye, “Hvorfor et fritt kraftmarked”, EFI’s User Conference on Power Production and Transmission, 7 June 1993 (in Norwegian)

[1.14] Walter Nicholson, “Microeconomic Theory, Basic Principles and Extensions”, Dryden Press International Edition, Fifth Edition, 1992

[1.15] Jean Tirole, “The Theory of Industrial Organization”, MIT Press, Ninth printing 1997

[1.16] Cody D. Walker, W. Timothy Lough, “A critical review of deregulated foreign electric utility markets”, *Energy Policy*, Vol. 25, No. 10, 1997, pp. 877-886

[1.17] John Casazza, “Reorganization of the UK Electric Supply Industry”, *IEEE Power Engineering Review*, July 1997, pp. 15-19

[1.18] P. Bornard, W. Kling, L. Martin Gomez, “Une analyse des conséquences techniques des changements d’organisation du secteur électrique”, Second Conference on the Development and Operation of Interconnected Power Systems “What limit for Interconnection?”, Budapest; 13-15 November 1996 (in French).

[1.19] “Effektreserver på nya elmarknaden”, Draft report of 1997-04-22 by EME Analys to the Swedish regulator NUTEK.

[1.20] Ivar Wangensteen, “Power Pooling Arangements. Review of International Best Practices”,TR F4746, SINTEF Energy Research, 1998.

[1.21] Several versions of “Electriciteitsplan” of NV Samenwerkende Electriciteits-Produktiebedrijven (SEP), prepared biannually until 1997 (in Dutch)

[1.22] I. Wangensteen, J.A. Holtan, “Organisering av energi/effektmarked”, EFI TR F4257, available as Enfo Report 87-1995 (in Norwegian)

[1.23] Lester H. Fink, Paul J.M. van Son, “On System Control Within a Restructured Industry”, *IEEE Transactions on Power Systems*, Vol. 13, No. 2, May 1998, pp. 611-616.

[1.24] “The Benefits and Deficiencies of Energy Sector Liberalisation”, Volume 1, World Energy Council, 1998

[1.25] “Composite Restated New England Power Pool Agreement”, Forty-Second Amendment, Attachment 2, Section 12 [cited 2000-10-30]. Available from Internet <http://www.iso-ne.com/main.html>

[1.26] “Amended and Restated Operating Agreement Of PJM Interconnection, L.L.C.”, Including FERC-Approved Revisions As Of September 18, 2000, Schedule 11, [cited 2000-10-30]. Available from Internet <http://www.pjm.com/>

[1.27] Bo Källstrand, “Spelreglerna leder till investeringar utomlands”, *Kraftordet*, No. 4, 1998 (in Swedish)

[1.28] Richard D. Tabors, “Reliability: Reality or the Power Engineers’ Last Gasp”, Proceedings of the 31st Annual Hawaii International Conference on System Sciences, January 6-9, 1998.

Chapter 2: THE CAPACITY BALANCE IN RESTRUCTURED POWER MARKETS

This chapter presents the capacity balance issue in restructured power markets from different viewpoints. The objective is to argue that shortage of peaking power and reserves may be expected in restructured systems in the long run. The argument will be made both on empirical and theoretical grounds.

It is important to notice at this point that the reference case for this chapter, and also for other chapters where relevant, is the "energy-only" market form. With this market organization, generators are paid only for the energy delivered. At this stage, payment of reserves is not considered, but the implicit assumption is that there is no *fixed* payment for reserves in the reference case, although there may be short-term, market based payment. (If there was a fixed payment, investment could always be made attractive by making the payment high enough, in which case one actually has chosen a particular policy to solve the problem.) The energy-only market is chosen as reference because it is the "natural" market form, normal in most other markets. One does not pay an additional fee on bread to guarantee baking capacity, and nobody requires airlines to keep enough planes available for all passengers that might wish to travel home the day before Christmas. So statements about the lacking viability of a power market to provide sufficient peaking power or reserves relate to the energy-only market. Other market organizations may be able to solve the problem, which is the theme of Chapters 6 and 7 of this thesis.

Section 2.1 discusses generation capacity, why it is needed and how it is supplied under various market settings. Section 2.2 is based on power system reliability theory, and its use in the determination of installed capacity. Section 2.3 discusses the role of uncertainty, and shows the effect of uncertainty with the help of a simple conceptual model. Section 2.4 compares the development of the capacity balance in Norway, Sweden and the UK. Finally, some conclusions are drawn in Section 2.5.

2.1 The capacity balance issue

The purpose of this Section is to provide the reason for why the balance between generation capacity and peak demand may become a problematic issue in restructured power systems. At this point, it is necessary to discuss the use of the word "balance". In a well-functioning market, there is a balance between supply and demand by definition. According to economic theory, the price will settle at a level where this balance is obtained. If demand temporary exceeds supply, the price will increase, depressing demand and/or increasing supply, until they are equal – and the other way round when supply exceeds demand. Electricity is special in that supply must equal demand at any point in time, because there are no storage possibilities. At the same time, it has been common to charge users fixed prices over longer time periods. It is clear that the combination of these two factors results in a market that cannot be well-functioning at all times, because there is no common market price seen by all participants that could have ensured a balance between supply and demand. Briefly, this is the background why there is a problem, and why it makes sense to discuss the capacity balance.

2.1.1 introduces some capacity definitions. 2.1.2 discusses the distinction between capacity- and energy-constrained power systems, and their development over time. 2.1.3 gives a brief overview over the expansion planning process in traditionally organized systems.

2.1.1 Capacity definitions

The IEEE Standard Dictionary of Electrical and Electronic Terms, defines capacity for electric-power devices as "The capacity of a machine, apparatus, or appliance is the maximum load of which it is capable under existing service conditions". This definition uses the notion of "load", which here will be interpreted as power, i.e. energy per time unit.

For a single generation unit, it seems straightforward to define this maximum power, given its existing service conditions (for example ambient temperature, current operating point). Even for a single unit, there may a number of reservations, which however, are less relevant for the current discussion. It is a lot more problematic to define the capacity of a system of generators. Because there is a certain outage probability for each unit and each plant, the total capacity is a stochastic number. Other concerns complicate matters utterly, e.g.:

- the grid may reduce the possibility to utilize all generators at full capacity, but to which degree will depend on which generators actually are online
- the possibility for grid outages increases uncertainty
- for run-of-river plants, availability of water may be difficult to predict

As an alternative to the capacity of the generation system, capacity may be viewed from the demand side. From this point of view, four definitions of capacity demand may be distinguished, as shown in Figure 2-1, which is based on Norwegian data from 1989-98 (Statnett). It shows the average per unit load duration curve together with the curve with the lowest load factor for the actual period.

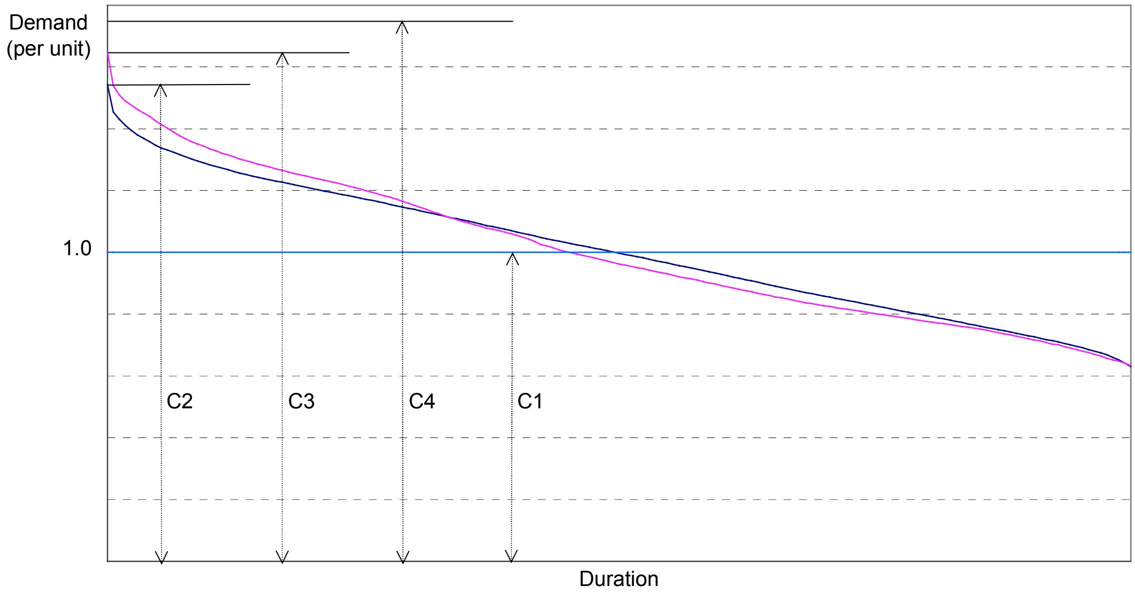

Figure 2-1: Illustration of concepts for demand for capacity

The figure represents the following concepts:

1. The capacity to satisfy the demand for energy during the year.
2. The capacity to provide normal peak load demand
3. The capacity to provide extreme peak load demand
4. The capacity to provide extreme peak load demand with agreed-upon reserve margins

The first form is mostly of theoretical interest, though it sets focus on the importance of the shape of the load-duration curve. The distinction between the second and third forms is important with respect to the uncertainty regarding peak demand even in the short run[1]. This uncertainty is due to annual variations in temperature. The impact of these variations depends on the share of the load that is used for either cooling or heating. The capacity to provide extreme peak load demand with agreed reserve margins must be available to secure reliable system operations under all circumstances, at least as long as price cannot be used effectively to constrain demand during peaking conditions.

2.1.2 Capacity- and energy constrained systems

One way to illustrate the role of generation capacity in power systems is to look at the distinction between energy- and capacity constrained systems. In the traditional single-owner view, a capacity constrained system is a system where the need for more capacity triggers

[1] This uncertainty is added to the long-term demand uncertainty, based on general economic development, fuel prices etc. A representation of historical maximum demand will always have a much more erratic character than the smoother development of energy consumption.

expansion. This is typically the case in thermal systems, where thermal capacity with a load factor of 80-85 % (7000-7500 hours) easily can provide the energy needed for demand with a load factor of 60-70 %. In an energy-constrained system, the need for more energy triggers expansion. This is the case in hydro-dominated system with sufficient storage capacity and for example a generation load factor of 45-50 % (4-5000 hours).

An illustration of the concepts is given in Figure 2-2. It shows optimal system expansion, given a known demand level at a certain stage. Fixed cost is assumed a linear function of system capability. If more capacity is built, fixed costs increase, while (net) operational costs decrease, due to the opportunity to utilize the system in a more flexible way and possibly the sales of surplus energy.

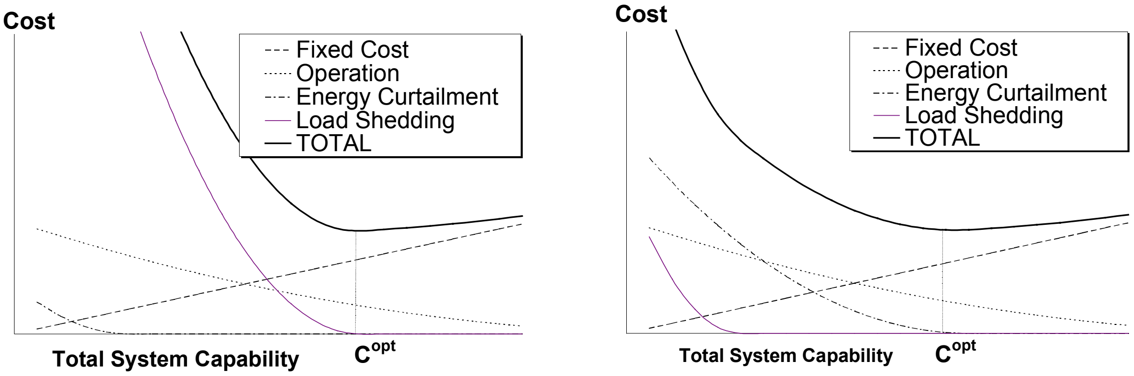

Figure 2-2: Optimal system expansion in capacity-constrained (left) and energy-constrained systems

Important entities in this context are “load-shedding” and “energy curtailment”. The former is here used for short-term, short-notice shedding of peak demand, the latter for longer-notice reduction of energy demand due to energy shortage. Both are forms for rationing, and are forms of market imperfections, resulting from the fact that consumers are insufficiently exposed to price variations that could adapt demand to supply. This problem is discussed in Chapter 3.

The left pane in Figure 2-2 shows a capacity-constrained system. The cost of load shedding is the dominating cost when the system expansion level is well below optimum. The cost of energy curtailment is zero far below the optimal level.

The right pane in Figure 2-2 shows an energy-constrained system. Here the cost of energy curtailment becomes the dominating cost below optimum, while the cost of load shedding is negligible at capacity levels of practical interest. The slope of the load-shedding cost curve is much steeper than of the energy-curtailment curve, reflecting its much higher cost[2].

[2] The Value of Lost Load in the UK system is for example GBP 2.50/kWh (USD 4.03/kWh) [2.2], and SEK 30.00/kWh (USD 3.56/kWh) in the Swedish system [2.4]. This may be compared with the cost of energy curtailment in Norway before restructuring of NOK 0.70-5.00/kWh (USD 0.09-0.64/kWh) (the highest number was only used for extreme and unrealistic levels of curtailment) and of SEK 3.00/kWh (USD 0.36/kWh) (Exchange rates summer 1999).

Energy surplus in capacity-constrained systems occurs because thermal plants have to meet peak demand, which gives a considerable slack at low-demand hours. Capacity surplus in energy-constrained systems may occur because the dimensioning criteria for hydro plants are not peak demand, but for example to minimize spill.

Two examples of cooperation between systems with different dimensioning characteristics, illustrating these features, are given in Appendix A.2 at the end of this chapter.

An interesting point is that this seemingly basic system characteristic is not necessarily a static property. A capacity-constrained system may develop in an energy-constrained direction because of load-factor increasing demand-side measures[3]. An energy-constrained system may move in the opposite direction because of load-factor reductions[4]. Shifts may also occur because of general changes in demand structure, for example changes in penetration of electrical heating or cooling. Related to the figures above, this means that the cost curves of load shedding and energy curtailment approach each other during such developments.

These tendencies, occurring isolated in separate systems, may be greatly enhanced when power exchange between systems with different characteristics increases. For the Norwegian system, stronger ties with neighbouring systems create the opportunity for large energy imports, relaxing the energy constraint. At the same time, export contracts have been signed, effectively decreasing the total demand load-factor. This means that the energy curtailment curve shifts to the left, while the load-shedding curve moves to the right. Eventually the last one may become the active constraint, turning the system to being capacity-constrained. The opposite development may occur in thermal systems that start importing peaking power from hydro-systems, shifting the load-shedding curve to the left, and the energy-curtailment curve to the right.

2.1.3 Expansion planning in traditional systems

In this Section, a short description will be given of general characteristics of the traditional planning process. The purpose is to give a background for comparison with the process in a restructured system.

Normally, some governmental body will have the supervisory responsibility for expansion of the power system. This organization basically takes responsibility for the following activities:

[3] This has happened in the Netherlands: its load-factor has been increased from 5800 hours in the seventies to 6200 hours in 1997 and a forecast of 6400 hours in 2006 [2.5]. The last number corresponds to 73 %. Although it would be far fetched to call the Dutch system energy constrained, 73 % is certainly approaching the 75 % settled down load factor for thermal plants used in IEA studies [2.6].

[4] This may be illustrated by the situation in Norway, where the load-factor has decreased from 6300 hours in the seventies to 5800 hours at the end of the nineties due to a fast growth of electrical heating after the oil price increases in the seventies. The development was probably influenced by the fact that partly capacity-based tariffs were substituted by pure energy-tariffs in the same period. Another reason for this development has been demand growth in domestic and commercial sector, while demand in the power intensive industry has been constant.

- prepare load forecasts (energy and/or capacity, cf. 2.1.2)
- identify relevant options to satisfy the demand forecast
- identify constraints
- compute least cost options that satisfy all constraints

In many ways, the most interesting point in this process is the satisfaction of "all constraints". At first, naturally there are numerous technical constraints regarding the various technologies, their viability in the system, reliability of supply, the extension and quality of the grid[5] etc. Second, society usually adds a number of non-technical constraints to the planning process. These may be various objectives that a society attempts to obtain by using energy policy as an instrument. Examples:

- the use of domestic fuels to decrease dependence on imports and preserve employment
- fuel diversification to avoid extreme dependence on one single energy source
- energy conservation and the use of renewables
- self-sufficiency, i.e. the objective to be dependent on domestic energy sources as far as possible
- resistance against nuclear power

Generally, the reduced opportunities to use energy policy as an instrument to obtain other objectives, is an important side effect of the restructuring process. Some view this as an advantage [2.3], while others are opposed to power sector restructuring exactly for this reason.

2.2 Power system reliability

Installed generation capacity is not 100 % available at all times. Firstly, planned outages are necessary for periodical maintenance. The length of the maintenance period varies with the type of plant: for thermal plants it may be from 4-6 weeks per year, for hydro normally just a few days, but with longer periods for major overhauls every 10-20 years. Under special conditions, annual maintenance of several weeks each year may also be necessary for hydro. Secondly, unplanned outages due to equipment failing reduce plant availability. Also this number is dependent on plant type, and has a stochastic character. For example in Holland, with a 100 % thermal system, unplanned unavailability varied between 5 and 12 % from 1971-87 [2.7], while the number for hydro plants in Norway is 1-2 %.

[5] Some authors argue that the lacking possibility to undertake integrated planning of generation and grid may be the major drawback of power system restructuring, an argument that cannot be neglected. On the other hand, in most developed countries there is considerable popular and political opposition against further extension of the high voltage transmission grids. This fact does to some extent change the grid operator's and owner's roles from planning grid expansion to maximizing the utilization of the existing grid.

The foreseeable character of planned maintenance does not constitute a planning challenge, but the resulting need for extra capacity has to be taken into account. However, the stochastic character of unplanned outages is a difficult planning problem, which has engaged researchers at least since the 1930's. The development of this research is described in [2.8] and [2.9].

Power system reliability is associated with the system's ability to provide a reliable supply of electrical energy. This very broad concept is usually subdivided in system adequacy and security [2.11]. Adequacy relates to the existence of sufficient facilities within the system to satisfy demand, which includes both generation, transmission and distribution facilities. Adequacy is associated with static conditions. Security relates to the ability of the system to respond to disturbances arising within that system. The majority of the research within the field is in the domain of adequacy assessment, but some work has been done on parts of the security problem, notably the quantification of operating capacity requirements.

The analysis of power system adequacy can be divided in three hierarchical levels. Hierarchical level I (HLI) is concerned only with generation facilities. Hierarchical level II (HLII) includes both generation and transmission and hierarchical level III (HLIII) also includes the distribution system in an assessment of consumer load point adequacy. In this section, the concern will be adequacy assessment at HLI only.

Two fundamentally different approaches have evolved over the years [2.12]. In North America the emphasis has been on the development of adequacy indexes for the total system and, at HLII, for every load point of the system. In Europe, particularly Italy and France, a Monte Carlo simulation approach has been used, where the "cost of reliability" is one of three cost components, the others being operating and capital costs.

2.2.1 The analytical approach : indexes

Historically, many utilities have measured the adequacy of both planned and installed capacity in terms of a percentage reserve. As pointed out in [2.8], there are a number of severe drawbacks with this measure: it does not recognize the very different requirements that may be needed in different systems, it does not consider load pattern changes over time and neither does it consider the impact of the addition of new large units. This last point is addressed by a modification, that demands that reserve requirements should equate the size of the largest unit plus a fixed percentage. Although this is an improvement, the economic aspects associated with different standards of reliability can be compared only by using probability techniques.

After some initial publications in 1933-34, a large group of papers was published in 1947, the most well known by Calabrese [2.12]. These papers introduced the methods that evolved in what now are known as "loss of load" and "frequency and duration" approaches.

The kernel of the analytical approach is a component state-model with two states: available and unavailable[6], which can be described by a single parameter, the forced outage rate. With this basic model, the probability of combinations of generator outages can be calculated, and

[6] More advanced models may have more states, e.g. derated states (with partial availability) or a four-state model for peaking units.

the cumulative probability of a specified level of MW being unavailable. To avoid excessive calculation times for large systems, outage combinations with a probability below a certain level are normally not considered (e.g. lower than $10^{-6} - 10^{-9}$).

This generation model can be convolved with a load model to produce system risk indexes. A number of different load models can be used, resulting in different indexes. In a simple and widely used model, each day is represented by its peak load. These can be arranged in descending order to form a daily peak load duration curve. Combined with a capacity outage probability table, the expected risk of loss of load, the Loss Of Load Expectation, *LOLE* can be calculated:

$$LOLE = \sum_{i=1}^{n} P_i(C_i - L_i) \quad \text{days/year} \qquad \textbf{(2-1)}$$

where

C_i	-	available capacity on day *i*
L_i	-	forecast peak load on day *i*
$P_i(C_i\text{-}L_i)$	-	probability of loss of load on day *i*
n	-	number of days considered

The result is the mathematical expectation of the number of days during which loss of load will be encountered. It is neither a frequency nor a duration. The simple basic model may be enhanced in many ways. Examples are dividing the year in several periods with varying availability due to maintenance, inclusion of derated states, or load forecast and forced outage uncertainty.

Another natural load model is the duration curve of the individual hourly values. With this model, the resulting index may also be regarded as a probability index, and is called Loss Of Load Probability or *LOLP*. Furthermore, the sum of the individual hours' capacity deficit may in this case be considered as a measure of expected energy curtailment or Expected Energy Not Served, *EENS* in MWh. By dividing *EENS* with expected energy demand *W* over the whole period a relative measure is obtained. By subtracting this figure from one, an Energy Index of Reliability (*EIR*) may be defined:

$$EIR = 1 - EENS/W \qquad \textbf{(2-2)}$$

These indexes do not give any indication of the frequency of occurrence of an insufficient capacity condition, nor its duration, which are the most useful indexes for customer or load point evaluation. Therefore, indexes representing frequency and duration measures offer increased compatibility in overall assessment. Naturally, this requires more generator outage data. In this case, transitions between the available and outage states are described by failure and repair rates, from which other quantities like mean time to failure/repair, cycle frequency and cycle time can be calculated [2.8]. The load may be represented by an individual state load model, where daily load is represented by a low and a high level, the latter equal to daily peak load. The number of hours with high load is called the exposure level. If this is one, the original *LOLE*-model reappears. Probability, frequency, duration and cycle time for loss of load events can be calculated with this model.

Apart from the use in HLIII analysis, it is not possible to state that the frequency and duration (F&D) method is better than the simpler *LOLE/LOLP/EENS* concepts, but it offers additional physical insight.

2.2.2 The simulation approach

In the simulation approach, the cost of reliability is obtained by developing a risk index in the form of the expected yearly curtailed energy, which is transformed into a cost using some appropriate value for unserved energy. The basic technique to accomplish this is the Monte Carlo method. The main advantage of the method is the feasibility of taking into account theoretically any random variable or contingency, but at the cost of excessive computing times. Moreover, the Monte Carlo method aims at total cost minimization, where outage costs are included in the total cost, while the index approach constrains the value of the relevant indexes. Provided that a sufficiently representative set of outages can be simulated and that the outage cost is reasonably realistic, the simulation approach should yield more theoretically correct results, because it is independent of somewhat arbitrary, engineering-judgement based index values.

2.2.3 The choice of reliability level

Regardless of which method and/or indexes are selected to determine the system adequacy, this does not answer the question of what the correct level should be[7]. The theoretical answer is simple: the level should be chosen such that the marginal cost of increasing it is equal to the marginal benefit to customers of this increase. In practice, the necessary calculations are well beyond today's level of knowledge (both methodical and especially on the necessary data) and computing power. In reality, the acceptable level is based on "engineering judgement", cf. [2.10]. Because of the "enormity of the (HLIII) problem" [2.11], there will always be uncertainty as to the validity of the answers. So especially if the theoretical result should indicate lower optimal reliability levels than engineering judgement, decision makers will be reluctant to change their policy, out of fear for the consequences.

According to the introduction to Part 1 of [2.9], "The ultimate purpose of assessing the adequacy of generation systems is to help in deciding how much additional capacity to install and when. There is very little published material concerning the reliability criteria used by utilities…". Some examples are:

[7] As stated at the end of the previous Section, the simulation approach avoids this problem – the reliability level will be a result of the cost minimization.

- In [2.14], criteria used by Canadian utilities are reported, based on a survey from 1979. The most used criterion was *LOLE*, with values of 0.1 and 0.2 days/year. One utility reported a value of 0.003 days/year when interconnections are taken into account.
- In Holland a more advanced version of the *LOLE* concept is used [2.7]. The year is divided in 13 four-week periods, and *LOLE* is calculated for each period. LOLE must not exceed 2 % in any period, and the geometric mean for all periods must not exceed 1 %. The criterion results in a loss of load frequency of 0.25/year and duration of 2 hours. The effect of interconnections is taken into account.
- In Sweden, a loss of load probability of less than 0.1 % has been used [2.15].

2.2.4 System adequacy after r estructuring

The previous Sections have given a rough overview over the methods and principal indexes used to assess generation adequacy in power systems. It appears that the theoretical correct level of adequacy, where its marginal cost equals its marginal benefit, is not used as a basis, because there are presently no methods available to calculate this optimum. Instead engineering judgement is used to establish index values, which traditionally have been handled as hard constraints in the capacity optimization process. In a restructured system, this approach comes under pressure, for two reasons:

- Parties involved may challenge the legitimacy of the values that are based on engineering judgement.
- It is not clear how the entity responsible for short-term security (normally the Independent System Operator or ISO) can enforce any specific level of long-term adequacy in an energy-only market.

The latter point naturally depends on the market organization. An analysis of several forms of market organization is given in Chapter 7. However, from the discussion so far, it is clear that it may be difficult or even impossible to sustain traditional levels of reliability in a restructured environment, at least without challenging the principles of the restructuring process.

2.3 Uncertainty

2.3.1 Sources of uncertainty

Uncertainty is an important factor in all planning. However, the economic risk facing utilities planning generation expansion in traditionally organized systems is limited[8]. The main uncertainties concern load growth and fuel prices, and the main objective is to satisfy demand

[8] This statement may need some moderation. Especially investment in nuclear technology has proven risky. Also at least one Norwegian utility faced severe economic problems after investment in expensive hydro plants in the early eighties. But such examples do not change the general rule.

"under all circumstances" at minimum cost. If demand forecasts turn out to have been too high, costs can be recovered through tariff modifications to (captive) consumers. If fuel prices increase more than forecasted, the effects can normally be compensated through consumer tariffs.

In centrally planned systems, a uniform discount rate is used. The main motivation for the choice of discount rate is to obtain a correct comparison between alternative projects. On this basis, the least cost projects can be chosen. Moreover, using the same discount rate for various public projects in principle assures an optimal allocation of resources between different sectors[9]. The choice of discount rate is not motivated by profitability of the company that realizes the project. Profitability is secured by calculating cost-covering tariffs[10].

In a restructured system, uncertainty is of direct concern for the results of profit-maximizing companies. When market prices substitute rates, it is impossible to compensate for the consequences of a wrong decision by modifying revenues.

From a general point of view, price is the only uncertain factor. More specific, a company assessing investment in new generation capacity will have to assess the important factors that affect electricity prices:

Structural factors:

- general economic development
- fuel prices
- demand growth by sector
- shifts in electricity utilization (more or less heating/cooling)
- new capacity built by incumbents
- new entrance

These factors are closely related to the development of the international economy.

Stochastic factors:

- temperature
- in hydro systems: inflow

In many countries, demand has some sensitivity for ambient temperature. Although this effect is smoothed out over a longer period, several (for the investor) unfavourable years may occur immediately after the investment has been made, threatening an otherwise positive result.

[9] At least this was the motivation for the choice of discount rate in the pre-restructuring Norwegian planning system.

[10] In the 1979 Energy Statement from the Norwegian Government the principle of pricing according to Long Run Marginal Cost was introduced. The principle was to some extent used in the next decade, and was one of the reasons for increasing electricity prices in Norway throughout the eighties.

Unpredictable factor:

- political decisions

The political conditions may change, depending on many different factors from opinions on emission control to budget balance. A potential investor has to take into account that the profitability of an investment may change overnight by unexpected political decisions.

In addition to these general uncertainties concerning all investments, additional uncertainty faces an investment in generation to provide peaking capacity. The need for marginal peaking capacity is the difference between two uncertain quantities: peak demand and available installed capacity. The resulting number is extremely volatile[11]. Another way to view this uncertainty is to look at the variability for the need for peaking power based on the load duration curve. Figure 2-3 illustrates this, where the three curves represent the upper parts of the lowest, expected and highest demand based on temperature variability[12].

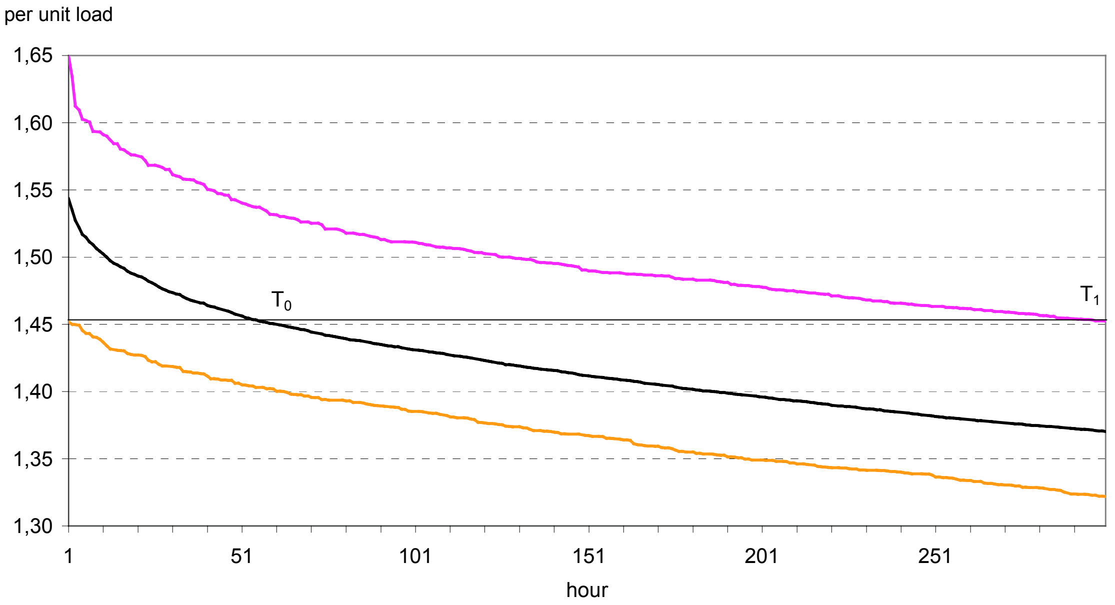

Figure 2-3: Average, highest and lowest per unit load duration curve (referred average annual load) for the Norwegian power system 1989-98 (source: Statnett)

Capacity exceeding 1.46 times average demand, is under normal conditions needed in T_0 (~50) hours. In a cold year, it is needed for T_1 (~300) hours, while in a warm year, it is not

[11] This can be illustrated by assuming that these quantities are normally distributed. For example in the Norwegian system, available capacity three years ahead may be a stochastic variable with an expected value with μ_p = 23 GW and σ_p = 2 % or 0.46 GW. Let demand be a stochastic variable with μ_d = 24 GW and σ_p = 2 % or 0.48 GW. Then the need for additional peaking capacity is a stochastic variable with $\mu_m = \mu_d - \mu_p$ = 1 GW and $\sigma_m = \sqrt{}\,(\sigma_d^2 + \sigma_p^2)$ = 0.66 GW or 66 %.

[12] The example is given for a system with electrical heating; in a system with cooling, the same reasoning is valid, but cold and warm must be exchanged.

needed at all. Because prices will vary also, the variability of the revenue from a peaking power plant will be even greater.

2.3.2 A simple price model

A simple, but interesting model to illustrate some principal relations was presented by Kay [2.16]: Let there be k techniques, each with a running cost a_i and capital cost b_i per demand cycle. Let there be n equal length subperiods, with x_{ij} per cycle the rate at which technique i is operating in period j, and x_i the capacity of technique i installed. Let q_j be demand in period j. The cost minimization problem may then be stated as:

$$\underset{x_i, x_{ij}}{\text{MIN}} \quad \sum_i \sum_j a_i x_{ij} + \sum_i b_i x_i \qquad \textbf{(2-3)}$$

subject to:

$$(u_{ij}) \qquad x_i - x_{ij} \geq 0 \qquad \forall i,j \qquad \textbf{(2-4)}$$

$$(v_j) \qquad \sum_i x_{ij} \geq q_j \qquad \forall j \qquad \textbf{(2-5)}^{13}$$

$$x_i,\ x_{ij} \geq 0 \qquad \forall\ i,j$$

Equation **(2-3)** describes the total operation and investment costs, **(2-4)** restricts generation from each technique to its capacity and **(2-5)** equals demand to generation. The dual variables corresponding to each constraint are shown in parentheses in front of the constraints. The dual formulation of the problem is:

$$\underset{v_j, u_{ij}}{\text{MAX}} \quad \sum_j q_j v_j \qquad \textbf{(2-6)}$$

subject to:

$$(x_{ij}) \qquad v_j - u_{ij} \leq a_i \qquad \forall i,j \qquad \textbf{(2-7)}$$

$$(x_i) \qquad \sum_j u_{ij} \leq b_i \qquad \forall i \qquad \textbf{(2-8)}$$

$$v_j,\ u_{ij} \geq 0 \qquad \forall\ i,j$$

The u's are interpreted as capacity charges and the v's as shadow prices of outputs. Using the complementary slackness properties of primal and dual, one obtains:

[13] The alternative formulation $\sum_i x_{ij} = q_j$ yields the same result because cost minimization results in the left-hand side equalling the right-hand side. However, the chosen formulation simplifies the interpretation of the dual, because with the equal sign, the v_j's would not be restricted to positive values.

from **(2-7)**:

$$v_j < a_i + u_{ij},\, x_{ij} = 0 \text{ or } v_j = a_i + u_{ij},\, x_{ij} > 0 \text{ or } v_j = a_i + u_{ij},\; x_{ij} = 0 \qquad \textbf{(2-9)}$$

from **(2-4)**:

$$x_{ij} < x_i,\, u_{ij} = 0 \qquad \text{or } x_{ij} = x_i,\, u_{ij} > 0 \qquad \text{or } x_{ij} = x_i,\; u_{ij} = 0 \qquad \textbf{(2-10)}$$

from **(2-8)**:

$$\sum_j u_{ij} = b_i,\, x_i > 0 \qquad \text{or } \sum_j u_{ij} < b_i,\, x_i = 0 \qquad \text{or } \sum_j u_{ij} = b_i,\; x_i = 0 \qquad \textbf{(2-11)}$$

For completeness, the last column represents degenerated cases, which have no practical impact on the conclusions. Equation **(2-9)** shows that if a technique is operating ($x_{ij} > 0$) the price v_j equals its running cost a_i plus the capacity charge u_{ij}. From **(2-10)** the capacity charge is zero as long as a technique is operating below its installed capacity. Thus the price in any period is equal to the running cost of any technique that is operating at a level less than its installed capacity x_i. All techniques with running cost greater than price will be inoperative **(2-9)**, and all for which running cost is less than price will be working at full capacity (**(2-9)** and **(2-10)**). Where all techniques operate at zero or capacity outputs, price is demand-determined at some level between the running costs of the most expensive technique operating and the cheapest technique not operating. Any employed technique will exactly earn its capacity cost from the capacity charges earned in the period when its running costs are below the prevailing price **(2-11)**. This can also be concluded directly from the fact that for this well-defined linear problem, the values of the expressions **(2-3)** and **(2-6)** are equal because of the strong duality property. Because **(2-3)** represents generator (fixed and variable) costs and **(2-6)** generator revenues, this proves that revenues cover costs.

If the periods are ordered according to decreasing demand, then the increase in demand between two consecutive periods will always be covered completely by one technique. As a result of this, the capacity of each employed technique will be equal to the difference in demand between two distinct demand periods. The interesting consequence of this is that the price v_j equals marginal cost of some employed technique in all periods, except those periods where the installed capacity of the marginal technique in that period is exactly utilized. With few technologies and many periods, this means that almost all technologies recover their capacity cost during periods that other techniques are price setting. The exception is the peak technology (with lowest b_i and highest a_i), that must recover its capacity cost in the single period it is used at its installed capacity, i.e. the only period that its $u_{ij} > 0$. A 3-period example in Appendix A.2.2 illustrates the concept in some more detail.

The fact that capacity costs are covered for all techniques is somewhat surprising, and indicates that there should be no peak-load problem because an investment in peak capacity fully recovers its costs. However, it is the result of two strong assumptions: planning is done under certainty and prices can be varied continuously. The effect of relaxing the first assumption is examined in the next Section.

2.3.3 The effect of uncertainty

It is possible to extend the simple model from the previous Section to include the effect of uncertainty. For this purpose, it is assumed that uncertainty is represented by a number of scenarios, each with a probability p_k. For each scenario, demand is assumed to be q_{jk}. In this case, the formulation changes to:

$$\underset{x_i, x_{ijk}}{\text{MIN}} \quad \sum_k \sum_i \sum_j p_k a_i x_{ijk} + \sum_i b_i x_i \qquad \textbf{(2-12)}$$

subject to:

$$(u_{ijk}) \quad x_i - x_{ijk} \geq 0 \qquad \forall i,j,k \qquad \textbf{(2-13)}$$

$$(v_{jk}) \quad \sum_i x_{ijk} \geq q_{jk} \qquad \forall j,k \qquad \textbf{(2-14)}$$

$$x_i,\ x_{ijk} \geq 0 \qquad \forall\ i, j,\ k$$

The meaning of the symbols is as before, with the difference that x_{ijk} here represents generation by technique i in period j in scenario k. The dual formulation of the problem now becomes:

$$\underset{v_{jk}, u_{ijk}}{\text{MAX}} \quad \sum_k \sum_j q_{jk} v_{jk} \qquad \textbf{(2-15)}$$

subject to:

$$(x_{ijk}) \quad v_{jk} - u_{ijk} \leq p_k a_i \qquad \forall i,j,k \qquad \textbf{(2-16)}$$

$$(x_i) \quad \sum_k \sum_j u_{ijk} \leq b_i \qquad \forall i \qquad \textbf{(2-17)}$$

$$v_{jk},\ u_{ijk} \geq 0 \qquad \forall\ i, j,\ k$$

In this case prices v_{jk} and u_{ijk} are scaled with the factors $1/p_k$, so $v'_{jk} = v_{jk}/p_k$ is interpreted as the shadow price for the output in scenario k[14]. Like before the complementary slackness properties of primal and dual can be used to analyse prices (the degenerated cases have been omitted here for clarity):

from **(2-16)**:

$$v'_{jk} < a_i + u'_{ijk},\ \ x_{ijk} = 0 \qquad \text{or} \quad v'_{jk} = a_i + u'_{ijk},\ \ x_{ijk} > 0 \qquad \textbf{(2-18)}$$

from **(2-13)**:

$$x_{ijk} < x_i,\ \ u'_{ijk} = 0 \qquad \text{or} \quad x_{ijk} = x_i,\ \ u'_{ijk} > 0 \qquad \textbf{(2-19)}$$

[14] It can be seen that this is correct by looking at equation **(2-15)**, that represents the total generator revenue. This can alternatively be written as $\sum_k \sum_j p_k v'_{jk} q_{jk}$, from which the relation between v_{jk} and v'_{jk} follows.

from **(2-17)**:

$$\sum_k \sum_j p_k u'_{ijk} = b_i, \quad x_i > 0 \qquad \text{or} \quad \sum_k \sum_j p_k u'_{ijk} < b_i, \quad x_i = 0 \qquad \textbf{(2-20)}$$

Prices equal the running costs of all units not running at their capacity limit. Also in this case capacity costs are recovered by the expected value of the capacity charges u_{ijk}. However, while expected total revenue of the generators *E(R)* is given by **(2-15)**, its variance is given by

$$Var(R) = \sum_k \frac{1}{p_k}\left(\sum_j q_{jk}{}^2 v_{jk}{}^2 \right) - E(R)^2$$

This uncertainty is especially harmful for peaking plants. The plant with the highest marginal cost a_i will have to cover its investment cost during a few short periods where all plants are running at their capacity limits. This is the only case where u_{ijk} for this plant is positive according to **(2-19)**. Without loss of generality, scenarios and periods can be defined such that this is the case for only one scenario *K* and one period *J*. In that case, the expected value of the revenue of this marginal peaking plant is $E(R_P) = q_{JK} v_{JK}$ and its coefficient of variation $\sigma(R_P)/E(R_P) = \sqrt{(1/p_K - 1)}$. For example for a plant that is on average needed every other year, $p_K = 0.5$, with a resulting coefficient of variation of 100 %. An example of the model including uncertainty is given in Appendix A.2.2.

2.3.4 Uncertainty and investment decisions

There is a wealth of economic literature about the effects of uncertainty on investment decisions, and it is well outside the scope of this thesis to probe this subject in any depth. Still it is appropriate to discuss shortly the concept of risk aversion. In economic theory it is common to model an individual's preferences by way of utility functions, cf. [2.18] Chapter 4. Put simply, a utility function describes how an individual's utility changes as a function of his wealth. A utility function is always increasing: the marginal utility of wealth is positive, people prefer more wealth over less. The shape of the function indicates the individual's risk attitude. If the utility function is linear, earning some amount gives the same numerical change in utility as losing the same amount, and the individual is called risk-neutral. If the utility function is convex, earning an amount gives a higher numerical change in utility than losing the same amount – the individual is risk-seeking. Finally, if the utility function is concave, earning an amount gives a lower numerical change in utility than losing the same amount. This is called risk-aversion, and it is generally assumed that this is the most common attitude towards risk.

A risk-averse individual (or company) will demand a compensation for taking a risk. As a result, a company will require a higher rate of return for the investment, or alternatively, use a higher discount rate to calculate its profitability. For investments in peaking capacity that have been shown to have great uncertainty, this means that very high discount rates will be used. Therefore, investment will only occur when extremely high prices are foreseen, and there are no risks of price caps.

2.4 Experience from existing markets

This Section will give a brief overview over the availability of capacity in the nineties in some countries. The objective is to find out if there is empirical evidence for the hypothesis that peaking capacity will become scarce in restructured power systems.

2.4.1 Norway

The Norwegian power system is almost entirely based on hydropower (99 % of installed capacity, the remainder mostly existing of CHP plants in industry and district heating). Installed capacity in 1999 is 28.4 GW, with an estimated normal annual production of 113 TWh. Gross total consumption in 1998 was 120.3 TWh, with a peak demand of 20.6 GW. Total consumption was covered with 116.7 TWh generation and 3.6 TWh import.

Net demand in Norway in 1998 existed of 71.5 TWh general demand (domestic, commercial, services, transport, agriculture) and 29.0 TWh energy intensive industry[15] (NVE). Moreover, 4.9 TWh was used by switchable boilers, and 0.8 TWh for seasonal[16] pumping. A considerable part of general demand is used for space heating.

Norway was one of the first countries to undertake a comprehensive restructuring. The current energy-only market is based on the Energy Act of June 1990, which in reality became effective with the introduction of the point tariff system in May 1992. A more comprehensive overview over the Norwegian restructuring process is given in [2.21] and [2.22]. Some production and demand data from 1991-1998 is given in Table 2-1.

[15] "Energy intensive industry" exists of smelters and various heavy metal industries, chemical industry and to a certain extent pulp and paper industry. Typical for this sector is a load factor between 90 and 100 %. The distinction between "general demand" and "energy intensive industry" has always been important in the Norwegian energy debate, because of the great share this industry takes of total electricity consumption, alleged subsidies, regional policy etc.

[16] Pumping in the Norwegian system has a seasonal character, in contrast with the daily pumping cycle used e.g. in the UK pumped storage plants. The purpose of seasonal pumping is to move water from small to large reservoirs in summer to save water for winter utilization.

Table 2-1: Norwegian power consumption and capacity demand and installation

	1991	1992	1993	1994	1995	1996	1997	1998
consumption (GWh)	99965	100441	101949	102926	104964	104146	105222	110666
installed capacity (MW)	27100	27100	27400	27500	28300	28400	28400	28400
maximum load (MW)	18700	18766	19418	19374	20302	21247	19528	20639
load factor (%)	66.1	66.2	66.0	66.6	65.4	61.1	67.6	66.7
maximum load with min load factor (MW)	20228	20329	20968	21134	21744	21247	21613	22541
reserve margin (%)	44.9	44.4	41.1	41.9	39.4	33.7	45.4	37.6
reserve margin with min load factor(%)	34.0	33.3	30.7	30.1	30.1	33.7	31.4	26.0

Because of the large share of space heating, Norwegian demand for peaking power is highly dependent on temperature. A simple way to correct for varying temperatures between years is to use the observed minimum load factor in the period concerned to calculate a "worst case[17]" peak demand. The reserve margin is define as:

$$margin = (P_{installed} - P_{max}) / P_{max}$$

where $P_{installed}$ equals installed capacity and P_{max} equals maximum demand.

Figure 2-4 shows the development of the reserve margins for Norway. The figure suggests that Norway still has a satisfactory reserve margin, although it is clearly decreasing, due to the fact that only minor investments in generation capacity have been made in recent years. Shutting down of hydropower units, in contrast with thermal units, cf. the next section, is normally not profitable: because of their low running costs, hydro units will over time always generate the electricity available from their inflow. However, when major refurbishment is necessary, the cost of this must be weighed against future revenues and costs. In this context, capacity in excess of 1000 MW may be mothballed in Norway in the period 2000-2002 [2.24].

The figure also shows a required reserve margin, which is of illustrative nature, because it is difficult to estimate and actually an improper measure of system adequacy, cf. Section 2.2.1. It is based on an estimated availability of hydropower of 89 %[18] and primary and

[17] I.e. worst case for the period concerned. Without further analysis, there is no way of saying how representative the corresponding temperature is in a longer perspective. This simple correction also oversees the fact that utilization of electricity may change over time, for example through an increasing or decreasing share of space heating.

[18] This is a controversial number, notoriously difficult to verify, partly because of its dependence on hydrological and hydraulic conditions. Investigations during the winter of 1996 indicated an availability of 87 and 84 % on two occasions. In this case, full utilization of remaining resources is assumed. This may prove difficult due to grid constraints.

secondary reserve requirements of 513 and 1000 MW respectively [2.23]. The reason that the required margin is somewhat decreasing is that the reserve requirement is an absolute number, which decreases as a fraction of peak demand.

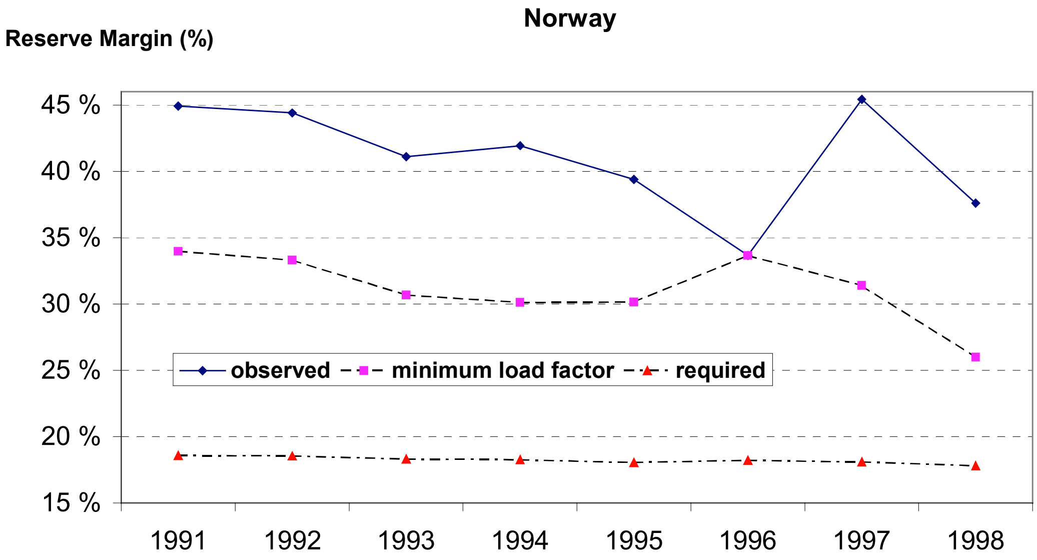

Figure 2-4: Development of reserve margins in the Norwegian power system

2.4.2 Sweden

Installed capacity in the Swedish system exists of 51 % hydro, 31 % nuclear, 17 % thermal (CHP, conventional steam and gas turbines) and 1 % wind, totalling 32.0 GW[19]. Energy production in 1998 amounted to 48 % hydro, 46 % nuclear, 6 % thermal and 0.2 % wind, totalling 154.2 TWh, of which 10.7 TWh was exported. Estimated normal annual hydro production is 64 TWh, which was exceeded with 9.7 TWh in 1998.

Net demand in Sweden in 1998 existed of 78.2 TWh for domestic use and services, and 54.6 industrial demand, of which 30 TWh can be characterized as “energy intensive”. From 1991 to 1995 between 4 and 8 TWh was used by switchable boilers, but no data are available for this from 1996. Also in Sweden a considerable part of general demand is used for space heating, though its share is lower than in Norway.

Sweden passed deregulation legislation in October 1995, and joined the existing Norwegian market structure in January 1996. A market operator, NordPool was formed, owned equally by Statnett and Svenska Kraftnät, the respective System Operators. More information is given in [2.21] and [2.22].

Some production and demand data from 1993-1998 is given in the table below.

[19] The statistical information in this Section is based on the annual publication “Kraftåret 19xx” from the Swedish Power Association (KvF), in Swedish with some English comments.

Table 2-2: Swedish power consumption and capacity demand and installation

	1993	1994	1995	1996	1997	1998
consumption (GWh)	131051	129675	132305	132446	132205	132800
installed capacity (MW)	34329	34532	34116	34251	34044	31994
maximum load (MW)	24400	24400	24400	26300	25000	24600
load factor (%)	65.6	64.9	66.6	61.9	65.1	66.6
maximum load with min load factor (MW)	25877	25586	26274	26300	26292	26476
reserve margin (%)	40.7	41.5	39.8	30.2	36.2	30.1
reserve margin with min load factor(%)	32.7	35.0	29.8	30.2	29.5	20.8

Figure 2-5 shows the development of the reserve margin for Sweden. This margin is obviously decreasing. Although it is difficult to establish absolute rules on a required margin, 20 % is obviously approaching this limit. Maximum load has been nearly constant for over a decade (the new record peak in 1998 was only 100 MW higher than the previous one in 1987, albeit at a considerably higher temperature in 1998). However, more than 2500 MW of conventional steamplant has been shut down, with the explicit motivation that it is unprofitable in a competitive market[20]. A report from the Swedish Energy Authority [2.25] acknowledged this situation, but claimed that sufficient import capacity will be available, a view that was heavily disputed by the industry.

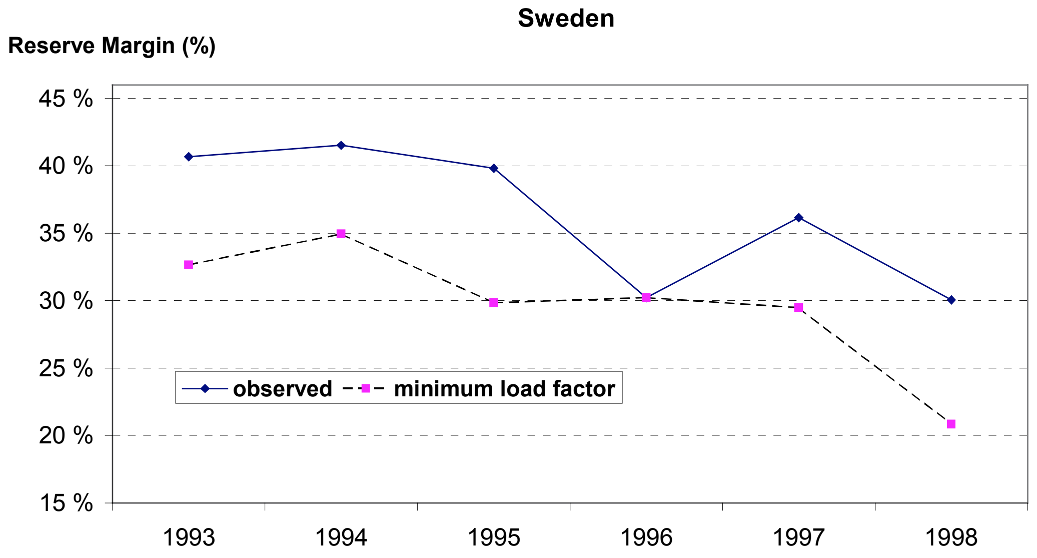

Figure 2-5: Development of reserve margins in the Swedish power system

[20] In addition, 600 MW of nuclear plant was shut down in 1999, but for altogether different (political) reasons.

2.4.3 England and Wales

Installed capacity in England and Wales at the end of 1996 existed of 57 % conventional steam plants, 18 % nuclear, 17 % CCGT, 6 % hydro (including pumped storage), 2 % gas turbines and oil engines and 1 % renewables[21] of a total of 73.3 MW. In terms of energy produced in 1996, the figures are 51 %, 27 %, 19 %, 0 %, 1 % and 1 % respectively of a total of 347.7 TWh. In addition imports were 16.7 TWh. Total consumption by final users in 1996 was 305.6 TWh, of which 103.1 was industrial demand and 202.5 domestic, commercial etc[22]. The share of space heating is low.

Often wrongly referred as the first open electricity market in the world, the reform that was effectuated on 1 April 1990, was still one of the earliest and probably the most debated and investigated. In the centre of this reform stands the Pool, to which all generation has to be sold, based on bids for 48 half-hour periods. In the present context, it is important to notice that the E&W market is not an energy-only market: it includes a capacity payment based on the half-hourly loss of load probability and the value of lost load (cf. the description in Section 7.2).

The supply side of the market originally existed of the three successors of the broken-up CEGB, National Power, PowerGen and Nuclear Power. Since then, several large new generators and many smaller IPPs have emerged. Retail competition was introduced gradually, and at present consumers, regardless of size can choose their supplier. Numberless descriptions of the system are available, among other a couple of the references already given.

Some production and demand data from 1991-1997 is given in the table below.

Table 2-3: UK power consumption and capacity demand and installation

	1991	1992	1993	1994	1995	1996	1997
consumption (GWh)	290840	291455	295142	291782	302637	314287	317486
installed capacity (MW)	70182	67499	69118	68998	70011	73261	72498
maximum load (MW)	54472	51663	54848	52362	55611	56815	56965
load factor (%)	66.9	70.1	66.7	70.8	68.4	69.6	69.9
maximum load with min load factor (MW)	54616	54329	54848	55589	57059	59274	59673
reserve margin (%)	28.8	30.7	26.0	31.8	25.9	28.9	27.3
reserve margin with min load factor(%)	28.5	24.2	26.0	24.1	22.7	23.6	21.5

[21] The statistical information in this Section is based on leaflets “UK Electricity” and “Electricity Industry Review” from the Business Information Centre of the Electricity Association.

[22] The leaflets use slightly different numbers under the headers “UK electricity supply and consumption” and “UK public distribution system”. Here the first numbers are used. Further on, the distinction between “energy intensive” and other industry is not made in these leaflets, and did not seem important enough in this context to pursue further.

Load factors have been calculated by dividing gross generation by maximum load, which gives slightly higher values than in the original EA source. Although the share of space heating is low, the load factor still varies considerably, which may partly be due to a limited use of electrical heating on very cold days. It is not clear if the higher load factors from 1994 are part of a trend or are caused by random variation.

Figure 2-6 shows the development of reserve margins for this system. There is a slight reduction in the observed margin. This is caused by a considerable layoff of conventional steam plant[23], which however is counteracted by a huge increase in independent CCGT plants. With the load factor of 1993, a considerable decrease in reserve capacity can be observed, despite the capacity element present in this market.

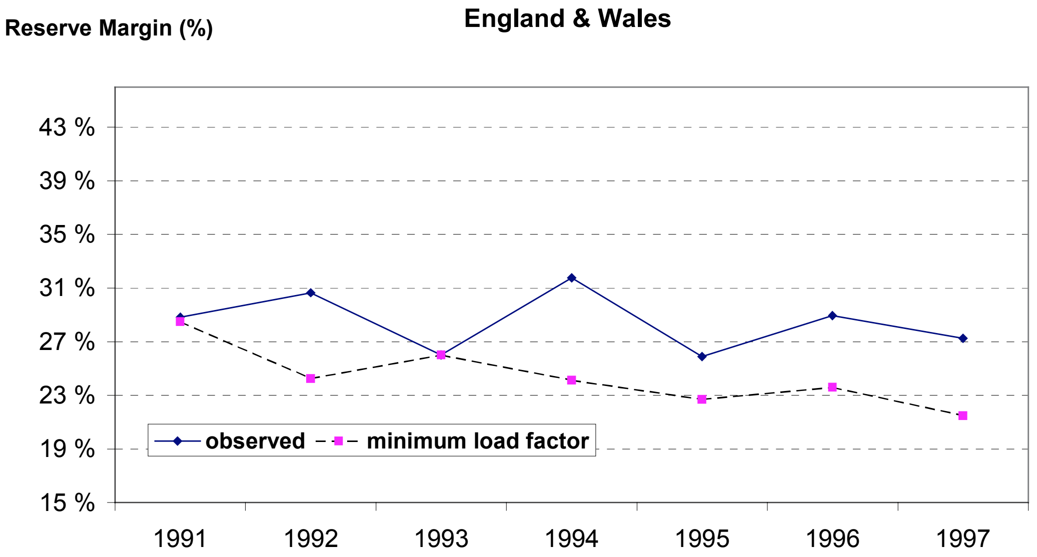

Figure 2-6: Development of reserve margins in the England and Wales power system

[23] The main reason is probably that conventional, coal based, plants cannot compete with modern natural-gas fired CCGT plants. The latter's superiority is partly due to low gas prices in the world markets and an accelerating technological development, but there are indications that distortions in the UK market also are a main reason for the "dash for gas". There have been discussions, if the still dominating generators National Power and PowerGen have deliberately closed existing capacity to be able to keep up market prices, cf. [2.29] and [2.1]: "Even though entry will cause the incumbents to set lower prices, considerable social loss is caused by the large and unnecessary induced investments in additional capacity." Another interesting development is illustrated by the Labour Government late-1997 temporary moratorium on new power stations (and not in the least by the following White Paper), seen by some as an attempt of support for the British coal industry. Apart from the question if this is the case or not, it puts focus on an important restructuring consequence, the greatly reduced opportunity for political control of the power sector. Even if this is often viewed as a positive effect, it also reduces the authorities possibility to pursue long-term beneficial objectives. For example may fuel diversification be a perfectly rational societal objective, which creates an externality. In this view, fuel diversification may be seen as an element of long-term security, not seen by the market, in the same way as the availability of reserves is an element of short-term security, not seen by the market either.

2.4.4 Other countries

In South-America, some electricity markets (Chile, Argentina) were restructured already in the 1980's, and consequently there is a long experience available in these countries. However, investment-stimulating elements in the form of payment for available capacity are included in these systems, so their experience does not answer the question if energy-only markets are viable.

The experience in the US is presently too short to draw any conclusions with respect to investment behaviour. The most developed markets are the PJM Interconnection and the New England and Californian systems. The two first have capacity obligations (discussed Chapter 7), which makes them unsuitable to answer the current question. As referred in Section 1.1, California has experienced capacity shortages, which partly may be due to generators' unwillingness to invest under the present uncertainty. But there are so many other factors influencing the situation that the Californian case does not provide strong evidence.

Another interesting example is Alberta in Canada, where the Power Pool of Alberta has been operating since 1 January 1996. This is also an energy-only market. There is clearly some concern regarding the market's ability to attract new capacity, cf. [2.26]: "The "supply cushion" within Alberta has been eroding over time." and [2.27]: "... there has been insufficient economic incentive to add generating capacity, despite supply constraints so severe that the system is often operating at or near full capacity".

Apart from the countries referred earlier, the experience in the other European countries and in Australia is too short to draw any conclusions.

2.5 Will the lights stay on?

Section 2.1 has introduced the central theme of this thesis: how to ensure that sufficient capacity will be available in restructured power markets. A better way to express this is: how to ensure that involuntary load shedding can be kept at a level that is acceptable to society. The distinction between these two formulations is essential. The former concentrates on the supply side, while the latter emphasizes a comprehensive view on both the supply and demand side of the market.

The objective of Sections 2.2, 2.3 and 2.4 has been to argue, from three different perspectives, that the supply of peaking power indeed is a problem in restructured systems. The first perspective is from traditional power system reliability theory. A short review of the theory has been given, and it has been argued that in a system with a pure energy spot market, the traditional levels of reliability cannot be sustained during peak load. Naturally, this will be possible by introducing additional instruments. This will be analysed later in this thesis.

The second perspective is through the effect of uncertainty. It is shown that investments in peaking capacity are uncertain, with the help of a simple conceptual linear optimization model for operation and investment. With the help of this model, it is shown that investment costs are covered on an expected value basis, but that the volatility of the revenue of the marginal peaking technology is extremely high. For this reason it is doubtful if profit-maximizing agents would invest in such peaking capacity.

The third perspective has been to investigate the existing markets in England, Norway and Sweden. In the two first, reserve margins are lower than they were in 1990, but not necessarily dramatically. This is somewhat difficult to judge because of climatic variability. For Sweden, the effect of restructuring on capacity has clearly been dramatic.

The arguments in this chapter show that in an energy-only market structure, the probability of involuntary load shedding may become unacceptably high.

2.6 References

[2.1] Richard J. Green, David M. Newberry, “Competition in the British Electricity Spot Market”, *Journal of Political Economy, Vol. 100, No. 5, 1992.*

[2.2] Frank A. Wolak, Robert H. Patrick, “The Impact of Market Rules and Market Structure on the Price Determination Process in the England and Wales Electricity Market”, UCEI Working Paper, February 1997 [cited 2000-10-30]. Available from Internet <http://www.ucei.berkeley.edu/ucei/PDFDown.html>

[2.3] Bernard Tenenbaum, Reinier Lock, Jim Barker, “Electricity Privatization. Structural, competitive and regulatory options”, *Energy Policy*, Vol. 20, No. 12, December 1992, pp. 1134-1160

[2.4] “Effektreserver på nya elmarknaden”, Draft report of 1997-04-22 by EME Analys to the Swedish regulator NUTEK.

[2.5] Several versions of “Electriciteitsplan” of NV Samenwerkende Electriciteits-Produktiebedrijven (SEP), prepared biannually until 1997 (in Dutch)

[2.6] “Projected Costs of Generating Electricity, Update 1998”, Nuclear Energy Agency, International Energy Agency.

[2.7] “Electriciteitsplan 1989-1998”, NV Samenwerkende Electriciteits-Produktie-bedrijven (SEP), 1989 (in Dutch)

[2.8] Roy Billinton, Ronald N. Allan, “Reliability Evaluation of Power Systems”, Pitman Publishing, 1984

[2.9] “Applied Reliability Assessment in Electric Power Systems”, Edited by Roy Billinton, Ronald N. Allan, Luigi Salvaderi, IEEE Press, 1991

[2.10] Richard D. Tabors, “Reliability: Reality or the Power Engineers’ Last Gasp”, Proceedings of the 31st Annual Hawaii International Conference on System Sciences, January 6-9, 1998.

[2.11] Roy Billinton, Ronald N. Allan, “Power-system reliability in perspective”, IEE Journal of Electron. Power, March 1984, Vol. 30, pp. 231-236, reprinted in [2.9]

[2.12] Luigi Salvadere, Roy Billinton, “A comparison between two fundamentally different approaches to composite system reliability evaluation”, *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-104, No. 12, December 1985, pp. 3486-3492, reprinted in [2.9]

[2.13] G. Calabrese, “Generating reserve capability determined by the probability method”, AIEE Transactions on Power Apparatus and Systems, Vol. 66, PP. 1439-1450, 1947, reprinted in [2.9]

[2.14] Roy Billinton, “Criteria used by Canadian utilities in the planning and operation of generation capacity”, *IEEE Transactions on Power Systems*, Vol. 3, No. 4, November 1988, pp. 1488-1493, reprinted in [2.9]

[2.15] “Från Monopol till konkurrens”, NUTEK 1991:6 (in Swedish)

[2.16] J.A. Kay, “Recent Contributions to the Theory of Marginal Cost Pricing: Some Comments” *The Economic Journal*, Vol. 81, No. 322, June 1971.

[2.17] S. Jonnavithula, R. Billinton, “Cost-Benefit analysis of generation additions in system planning”, *IEE Proceedings-Generation Transmission and Distribution*, Vol. 145, No. 3, May 1998, pp. 288-292

[2.18] Thomas E. Copeland, J. Fred Weston, “Financial Theory and Corporate Policy”, Third Edition, Addison-Wesley Publishing Company, 1992

[2.19] Shimon Awerbuch, “The Surprising Role of Risk in Utility Integrated Resource Planning”, *The Electricity Journal*, April 1993.

[2.20] Avinash K. Dixit, Robert S. Pindyck, “Investment under Uncertainty”, Princeton University Press, 1994.

[2.21] Richard D. Christie, Ivar Wangensteen, “The Energy Market in Norway and Sweden: Introduction”, Power Engineering Letters in *IEEE Power Engineering Review*, February 1998, Vol. 18, No. 2.

[2.22] Richard D. Christie, Ivar Wangensteen, “The Energy Market in Norway and Sweden: The Spot and Future Markets”, Power Engineering Letters in *IEEE Power Engineering Review*, March 1998, Vol. 18, No. 3.

[2.23] B.H. Bakken, “Technical and Economic Aspects of operation of Thermal and Hydro Power Systems”, *PhD thesis* NTNU, Trondheim, Norway, 1997

[2.24] “The long-term capacity supply curve”, ECON-notat 35/2000, Prosjekt nr. 32630, KLM/JBr/LiP, March 2000 (in Norwegian)

[2.25] “Uppdrag att kartlägga och analysera tilgång och behov vad gäller reservkraftproduksjon i det svenska elsystemet”, 15 January 1999, Dnr: 02-98-3136, Energimyndigheten (in Swedish)

[2.26] “Market Surveillance Administrator Report on Power Pool of Alberta Prices – Summer 2000”, [cited 2000-10-27]. Available on Internet <http://www.resdev.gov.ab.ca/electric/>

[2.27] Andrew J. Roman, “Legal Responsibility for Reliability in the New Competitive Electricity Markets in Canada”, in Reliability in the New Market Structure (Part 2), *IEEE Power Engineering Review*, January 2000, Vol. 20, No. 1.

[2.28] “Conclusions of the Review of Energy Sources for Power generation and Government response to fourth and fifth Reports of the Trade and Industry Committee”, Whitepaper from the Department of Trade and Industry, October 1998

[2.29] Richard Green, “England and Wales – A Competitive Electricity Market?”, paper presented at the Nordic Energy and Society Programme, Oslo, October 1998.

Appendix A.2.1: Cooperation between hydro- and thermal systems

A thermal generation system is always capacity constrained, whereas a hydro system can be either capacity constrained or energy constrained. In any case, there will be an economic benefit in linking the two systems together.

1) Capacity constrained hydro system

This means that there is water available to increase the power generation in off-peak periods. Demand is variable, and the system is designed to cover peak demand. There may be load shedding during peak demand. In off-peak periods, there is spare capacity and water available. In many cases a hydro generation system in an early stage of development will be capacity constrained.

The benefit of connecting such a hydro system to a thermal one is an increased hydro generation in off-peak periods at almost zero cost. That will replace thermal generation and thereby save fuel in the thermal system. The net effect is a reduced operation cost in the total system. A schematic illustration is shown in Figure A2-1.

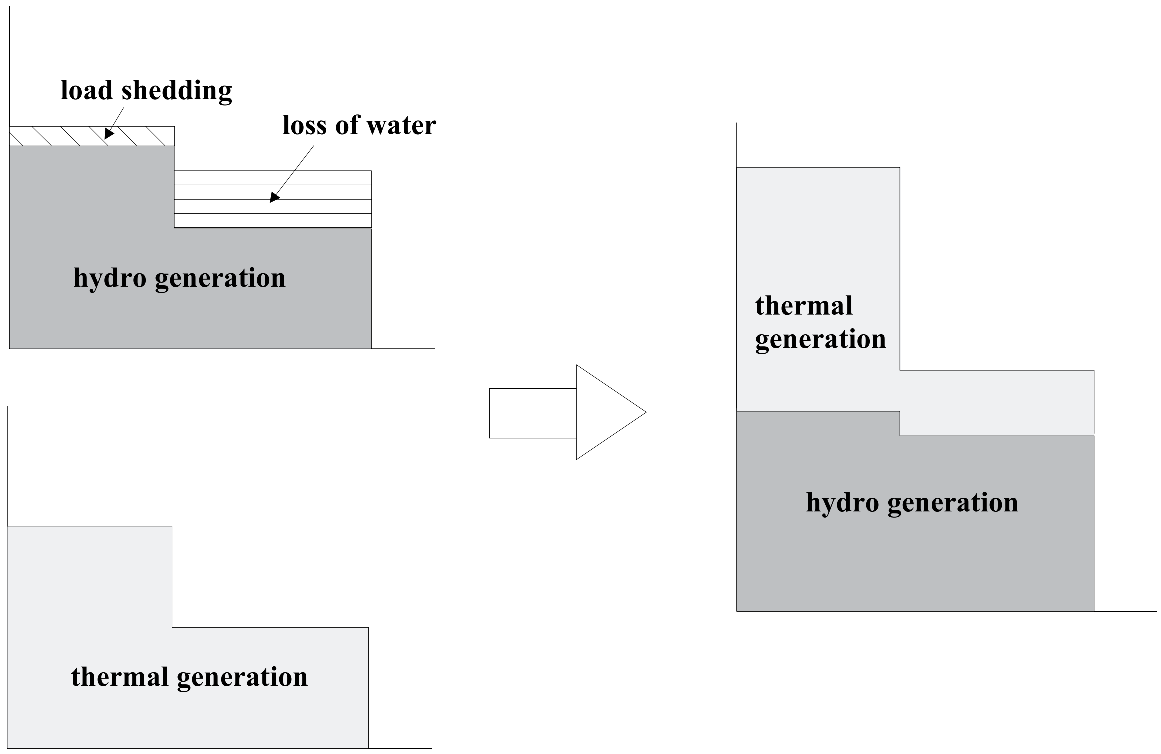

Figure A2-1: Illustration of the benefit of connecting a capacity constrained hydro system with a thermal system

The two panels to the left show the situation in each system separately. The hydro system suffers load shedding during peak, but has overflow during off-peak. Combining the systems relieves the load shedding, and makes it possible to utilize the excess generation capacity of the hydro system during low load conditions. Naturally, this requires that the thermal system has ample capacity, and that it is technically possible to operate with great load variations. The benefit for the hydro system comes from the relieve of load shedding, and for the thermal system from reduced total energy generation.

2) Energy constrained hydro system

The basic concern in an energy-constrained system, is the amount of water available. An energy constrained hydro system will normally have a surplus of capacity. Figure A2-2 illustrates cooperation between two systems with the same demand curve as in the previous example but in this case, the hydro system is energy-constrained. There is no load-shedding, and no overflow during off-peak. After interconnection, the hydro system provides peaking capacity, while the thermal system generates energy for the hydro system during off-peak, resulting in a flatter, more economic production profile.

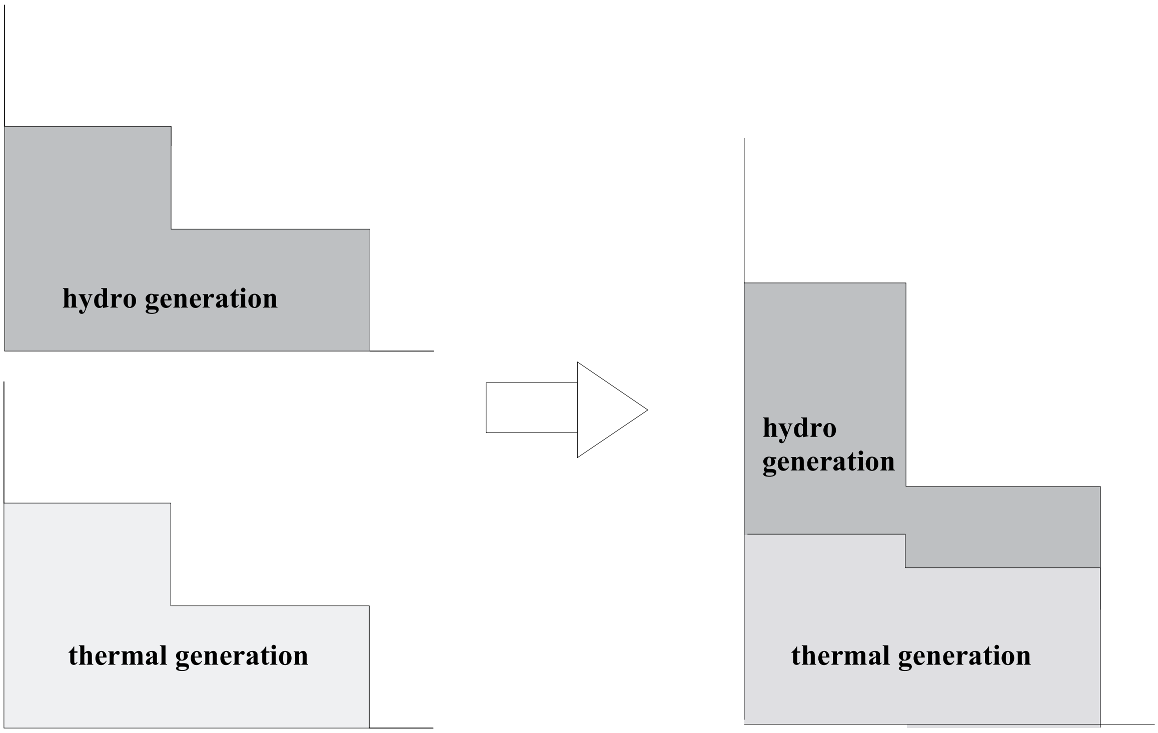

Figure A2-2: Illustration of the benefit of connecting an energy constrained hydro system with a thermal system

APPENDIX A.2.2: EXAMPLES OF A SIMPLE PRICE MODEL

The model described in Section 2.3.2 and 2.3.3 can be illustrated with a 3 period – 2 technology example. Demand for each period is shown in Figure A2-3.

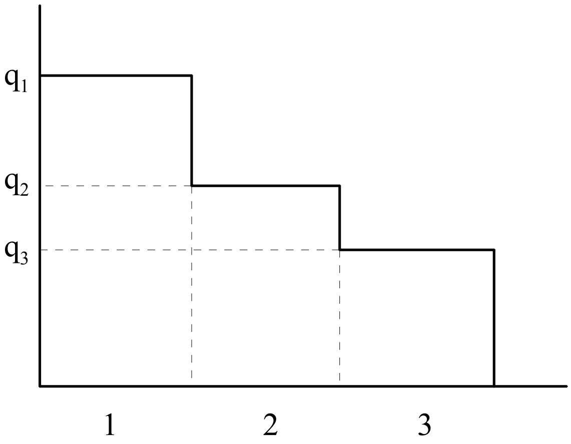

Figure A2-3: Simple 3-period demand model

Technology 1 is supposed to have the lowest capital cost and the highest running cost. Demand is sorted in descending order. Because this is an LP problem, there will always be a solution lying at a vertex point. For the present problem this means that in the optimal solution and starting with lowest demand, a technology will always either completely cover incremental demand in a period, or not be used at all in that period. Disregarding the solutions where one technology covers demand in all periods (because the other has either too high running or too high capital cost), there are two possible solutions to the 3-period – two technology problem:

A: technology 1 is used only in period 1, i.e. it covers the difference in demand between periods 1 and 2; capacity of technology 2 equals demand in period 2.

B: technology 1 is used in periods 1 and 2, i.e. it covers the difference in demand between periods 1 and 3; capacity of technology 2 equals demand in period 1.

For case A, the following equations emerge:

$$v_1 = a_1 + u_{11}$$
$$v_1 = a_2 + u_{21}$$
$$v_2 \le a_1$$
$$v_2 = a_2 + u_{22}$$
$$v_3 \le a_1$$
$$v_3 = a_2$$

Now $u_{11} = b_1$ and $u_{21} + u_{22} = b_2$. Technology 1 fully recovers its capital cost during the "peak" period 1, while technology 2 covers its capital costs partly in period 1 and partly in period 2. $u_{12} = u_{13} = u_{23} = 0$, $u_{21} = a_1 - a_2 + b_1$ and $u_{22} = b_2 - b_1 + a_2 - a_1$, which is positive because of the assumption about technology.

For case B, the solution becomes:

$$v_1 = a_1 + u_{11}$$
$$v_1 = a_2 + u_{21}$$
$$v_2 = a_1$$
$$v_2 = a_2 + u_{22}$$
$$v_3 \le a_1$$
$$v_3 = a_2 + u_{23}$$

In this case, $u_{11} = b_1$ and $u_{21} + u_{22} + u_{23} = b_2$. Technology 1 recovers its capital costs during period 1, while technology 2 recovers its capital costs during all 3 periods. $u_{12} = u_{13} = 0$, $u_{21} = b_1 + a_1 - a_2$, $u_{22} = a_1 - a_2$ and $u_{23} = b_2 - b_1 - 2(a_1 - a_2)$.

A slightly more elaborate 5-period – 3-technnology example is used to illustrate the model of Section 2.3.3 with some numerical results. Demand scenarios and their probabilities are given in Figure A2-4:

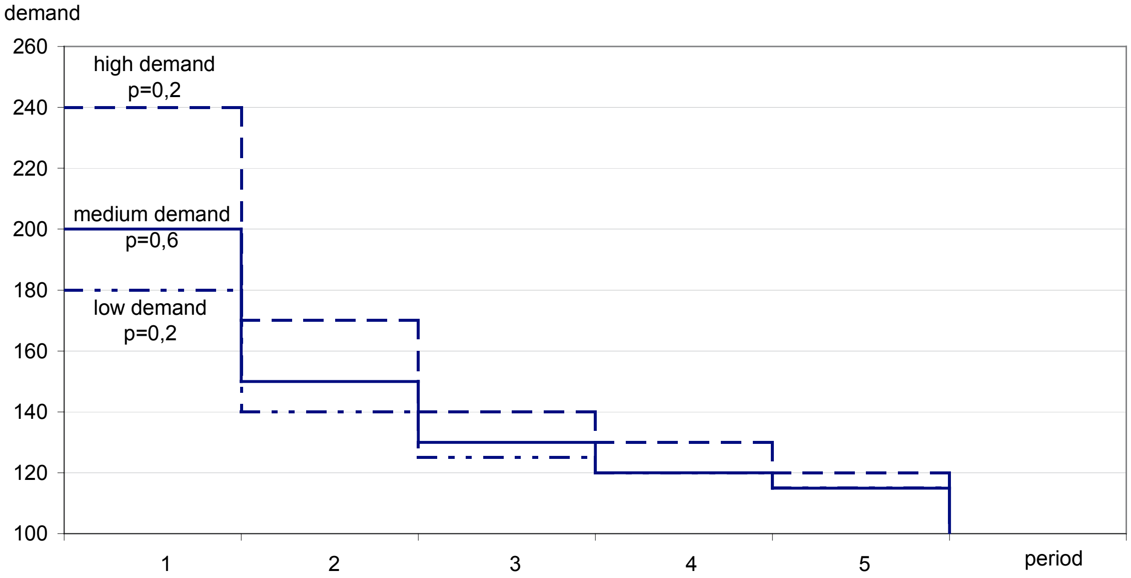

Figure A2-4: Example demand scenarios

For the three technologies the following costs are given:

Table A2-4: Example model costs

technology	variable cost (a)	capital cost (b)
1	12	5
2	5	10
3	2	18

Capital costs are artificially low in this example because of the short number of periods. The model was solved with AMPL using CPLEX as solver. Optimal installed capacities are given by $x_1 = 40$, $x_2 = 70$, $x_3 = 130$. The other results are given in the following table, where the meaning of the symbols is as in Section 2.3.3. The last column shows the technology that is directly price setting through its marginal cost. Where no value is given, price is between the marginal cost of two technologies, except for the highest load period, where price is such that the capital cost of technology 1 is covered.

Table A2-5: Example model results

scenario / prob.	period	demand	tech. 1 x_{1jk}	tech. 2 x_{2jk}	tech. 3 x_{1jk}	u'_{1jk}	u'_{2jk}	u'_{3jk}	v'_{jk}	price setting
high / 0.2	1	240	40	70	130	25.0	32.0	35.0	37.0	-
	2	170	0	40	130	0.0	0.0	3.0	5.0	2
	3	140	0	10	130	0.0	0.0	3.0	5.0	2
	4	130	0	0	130	0.0	0.0	3.0	5.0	-
	5	120	0	0	120	0.0	0.0	0.0	2.0	3
medium / 0.6	1	200	0	70	130	0.0	6.0	9.0	11.0	-
	2	150	0	20	130	0.0	0.0	3.0	5.0	2
	3	130	0	0	130	0.0	0.0	1.3	3.3	-
	4	120	0	0	120	0.0	0.0	0.0	2.0	3
	5	115	0	0	115	0.0	0.0	0.0	2.0	3
low / 0.2	1	180	0	50	130	0.0	0.0	3.0	5.0	2
	2	140	0	10	130	0.0	0.0	3.0	5.0	2
	3	125	0	0	125	0.0	0.0	0.0	2.0	3
	4	120	0	0	120	0.0	0.0	0.0	2.0	3
	5	115	0	0	115	0.0	0.0	0.0	2.0	3

The numerical results confirm the conceptual arguments. In period 1 of the "high" scenario, all technologies produce at their installed capacity, and the price is such that technology 1 covers its capacity cost. In period 1 of the "medium" scenario, the price is between the marginal costs of technologies 1 and 2, and in period 3 of the "medium" scenario, it is between the marginal costs of technologies 2 and 3. Technology 1 exactly covers its capacity cost in the high load period of scenario "high": $0.2 \cdot 25 = 5$. Technology 2 covers its capital cost in period 1 of scenarios “high” and “medium”: 0.2 · 32 + 0.6 · 6 = 10. Technology 3 covers its capital cost during several periods and scenarios. The coefficient of variation of the revenues of technology 1 are given by:

$$\frac{\sigma_p}{E(R_p)} = \frac{\sqrt{0.2 \cdot 37^2 - 7.4^2}}{7.4} = \frac{14.8}{7.4} = 2$$

which conforms with Section 2.3.3 because $\sqrt{1/0.2 - 1} = 2$.

Chapter 3: PRICING THEORY AND CONSUMER PREFERENCES

This chapter develops a theoretical foundation for the analysis of the relations between pricing of, demand for and investment in capacity. The chapter starts with a review of a selection of the literature on the subject of peak-load pricing in traditionally regulated systems. Although written for a different environment, several of the insights from this literature are valuable also after restructuring. A major problem, assessed directly or indirectly by most authors, is how to combine the benefits of prices based on short-run marginal costs with long-run revenue reconciliation. In some sense, this is also the central challenge for the satisfaction of peak-demand in restructured power systems. Section 3.2 discusses the role of the generation system in the provision of reliability. It may be regarded as a continuation of Section 2.3, and at the same time a preparation for Chapter 4. The last part of the current chapter is concerned with the theoretical foundation of an active demand side assessment. The basic idea is that present market organization obscures the underlying demand elasticity. The welfare gains of a more flexible demand side handling are discussed. Section 3.4 deals with an alternative model, where electricity consumption and reliability are viewed as two different products, where consumers receive utility from both. This gives insight in how electricity as a product might be tailored to consumer preferences.

3.1 A review of electricity pr icing theory

3.1.1 Historical overview

Optimal pricing of electricity has been a subject of research for several decades, and a huge literature is available. This section will give a brief overview over some important contributions to this literature, focussing on items relevant for pricing and availability of peak-load capacity. A broad survey of the theory is given in [3.1].

A natural starting point for a discussion is a paper by Boiteux, first published in 1949 in French [3.2] and later in English in 1960 [3.3]. The analysis of Boiteux is static: it considers pricing and optimal investment for a given demand level. In the first part of the paper, Boiteux discusses the situation with constant demand. His initial conclusion is that "The theory of sale at marginal cost is concerned with the short-term marginal cost[1] (…) for existing plant". However, it is shown that this will lead to prices too low to pay for investment costs. It is so proved that optimal pricing implies a price equal to short-run marginal cost, which, provided there is an optimal investment policy, equals long-run marginal cost, cf. Figure 3-1.

[1] Boiteux uses the term "differential cost"

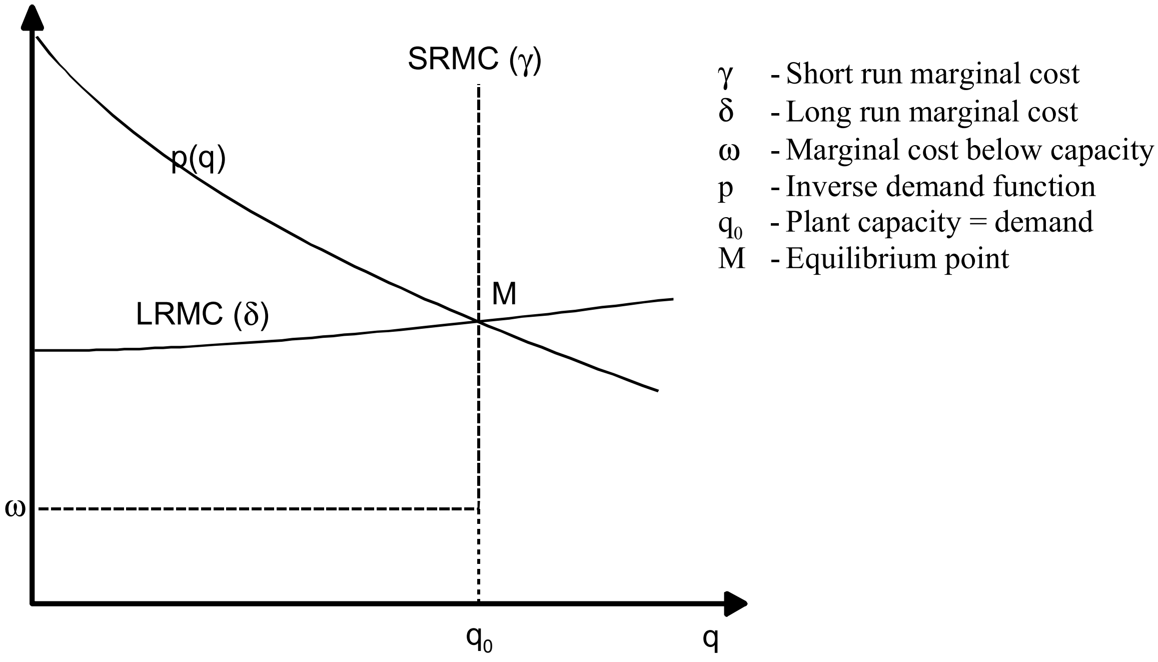
SRMC (γ)
p(q)
M
LRMC (δ)
ω
q₀
q
γ - Short run marginal cost
δ - Long run marginal cost
ω - Marginal cost below capacity
p - Inverse demand function
q₀ - Plant capacity = demand
M - Equilibrium point

Figure 3-1: Determinations of the optimum scale for plant of rigid capacity [3.3]

Next, Boiteux moves to the case of periodical loads, first exemplified with two periods. He then shows that, in an optimal situation, the price during the low load period must be equal to short-run marginal cost. In the high load period, the price is set equal to short-run marginal cost plus the plant development cost: “Each demand must pay for its relevant energy cost. The peak demand bears the complete plant charge 2π, as power cost”, cf. Figure 3-2, where the factor 2 comes from the fact that two periods of equal length have been assumed.

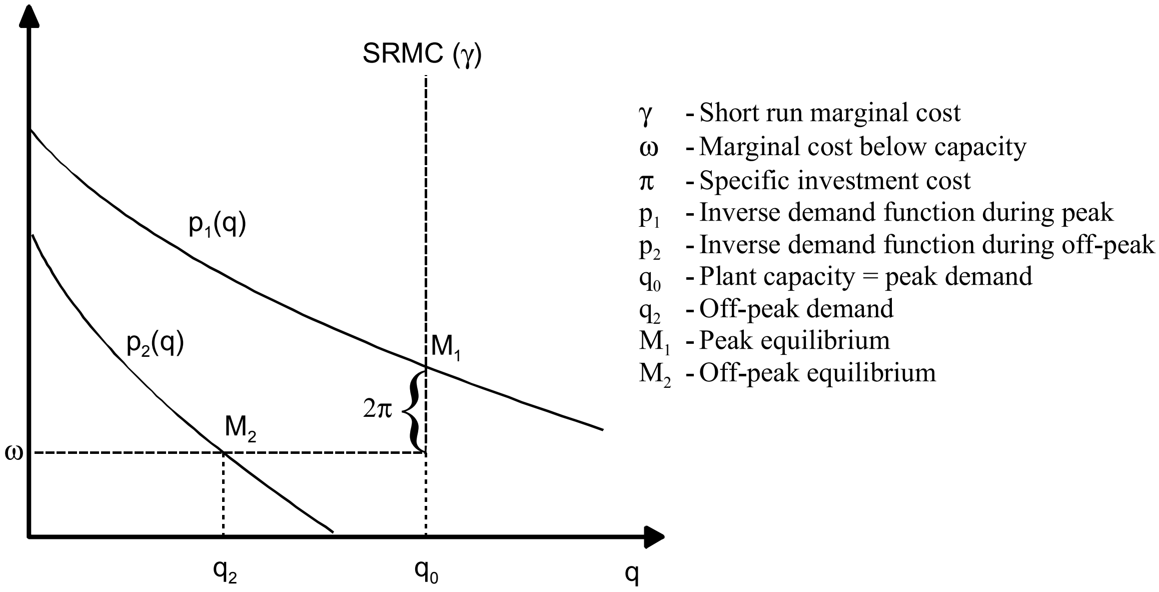
SRMC (γ)
p₁(q)
p₂(q)
M₁
2π
M₂
ω
q₂
q₀
q
γ - Short run marginal cost
ω - Marginal cost below capacity
π - Specific investment cost
p₁ - Inverse demand function during peak
p₂ - Inverse demand function during off-peak
q₀ - Plant capacity = peak demand
q₂ - Off-peak demand
M₁ - Peak equilibrium
M₂ - Off-peak equilibrium

Figure 3-2: Pricing for the two-period case [3.3]

Boiteux makes two essential assumptions:

- Plant capacity can be adapted continually and quickly up and down to match demand perfectly.
- Tariffs must be stable.

A paper by Williamson [3.4] in 1966 offered several new contributions. One was the development of an over-all “effective demand for capacity” relation, which directly provided optimal capacity in a straightforward graphical way. An important insight not treated satisfactory by many other authors, was the effect of indivisibility of new capacity. Together with uncertainty and a building lag, the result of this is that in the short run, capacity will seldom be optimal. In this case the optimal price might be lower or higher than the classical $b+\beta$ (or $\omega+2\pi$ using Boiteux’s symbols).

The models of Ralph Turvey (e.g. [3.5],[3.6]) have a more dynamic character, because an investment’s impact of future system operational costs is taken explicitly into account. Turvey develops a short-run pricing rule and an investment rule:

- “The price of electricity at each time, place and voltage is set equal to marginal running cost per kWh delivered or, if it be higher, at *the level necessary to restrict demand to capacity*”
- “An investment is undertaken if the present worth to consumers of the consequential change in their supplies, less the present worth of the total system costs with and without the investment is positive”

The short-run pricing rule is, as stated by Turvey, a direct expression of the analysis of Boiteux, as illustrated in Figure 3-2. The justification for this interpretation of short-run marginal cost is “simply the judgement (…) that rationing by price is preferred to rationing by power cuts”. Both rules require the consumer to pay the incremental running costs PV[dU·m], where PV stands for present value, dU is the load increment and m is the marginal cost[2]. In addition, the long-run rule requires him to pay dP (the incremental capacity demand) times the present worth of incremental capacity costs. The short-run rule requires him to pay dP times the excess of price over incremental running cost per kWh which is necessary to restrict demand to equal capacity at times of peak. Turvey shows that, if the optimal investment rule is followed (under the assumption of certainty in demand forecasts), the long-run and short-run rules result in the same price.

An important but controversial contribution was given by Brown and Johnson [3.7]. They were the first to model uncertainty explicitly, and came to the surprising conclusion that the optimal price should equal marginal cost. Naturally, others questioned this conclusion, which

[2] It should be noted that PV[dU·m] is not equal to the marginal cost at any given point in time, and consequently, does not imply that marginal consumer utility (or willingness-to-pay) equals marginal cost. In other words, it does not result in an optimal short-run solution.

appeared to depend on the crucial assumption that demand could be rationed costless according to willingness to pay. Visscher [3.8] showed that different assumptions regarding rationing policy imply different results. However, even with the bluntest instrument, random rationing, he still demonstrates an optimal price lower than Boiteux's and Turvey's classical b+β.

Carlton [3.9] uses a multiplicative (in contrast to Brown and Johnson's additive) model to describe demand uncertainty, and shows that in this case the optimal price exceeds b+β if random rationing is used. Furthermore, he makes the important observation with respect to rationing that: "*When goods are not always available to consumers, the probability of obtaining the good becomes a characteristic of the good*". On this background, he challenges the conventional use of expected surplus as the welfare function. A more comprehensive treatment of this view is given in [3.10]. This is an interesting basis for the analyses in Sections 3.3 and 3.4.

Meyer [3.11] presents a general treatment of uncertainty in a multi-period model. He confirms the results of Brown and Johnson as a special case, but acknowledges the dependence on the rationing strategy for the utility maximizer. For the profit maximizer, however, he concludes that the rationing strategy is irrelevant, because he is only concerned with his own costs and revenues. He criticizes the use of the welfare criterion because of the important role of rationing, which "(…) requires obtaining an estimate of *marginal* demand schedules, i.e. schedules of marginal willingness to pay. (…) the prospects for empirically determining it seem somewhat less than sparkling." This is an argument for organizing the market in such a way that consumers can reveal their real willingness to pay. Meyer justifies "*apparent* excessive investment" by stating that this "may be simply an optimal economic decision when the firm must meet high reliability standards for the service it provides". This may true for the firm, but one can question the economic basis for this level of security.

The book of Crew and Kleindorfer [3.12] gives a broad treatment of the area of pricing and investment. In their assessment of multi-period pricing investment with diverse technology, the authors develop a model that shows the strong interaction between optimal technology and optimal pricing. The more diverse the available technology, the less will be the tendency to flatten demand through peak-loading, i.e. the flatter the optimal price schedule will be. They also point out that inefficient rationing leads to high price differences between peak- and off-peak periods. Price, capacity and reliability are higher the less efficient rationing is. Crew and Kleindorfer state that reliability constraints (for example like LOLP < 0.1 days/year) may be thought of as a surrogate for rationing costs. Highly inefficient rationing requires very high reliability at optimum, and the problem then reduces to an equivalent deterministic problem.

3.1.2 The controversy about LRMC and SRMC

The discussion in the previous section makes clear that there has been considerable disagreement about correct pricing and optimal capacity investment in traditional regulated environments. On the one hand stand the proponents of the basic and intuitively appealing ideas of Boiteux, where price in off-peak periods is set to SRMC, and in peak-periods to

SRMC + β (=LRMC), where β is the appropriate share of investment costs to ensure cost recovery. On the other hand stand the authors who, on various grounds, argue for peak-load prices lower than SRMC + β, or even down to SRMC. Some of the authors argue that there is no contradiction here, because under optimality SRMC=LRMC.

In [3.14], Andersson and Bohman contest this view. They point out that this conclusion is critically dependent on the unrealistic assumption that capacity is continuously variable both when expanding and contracting, and that the investment policy follows a correct forecasting. As stated in [3.14], it has been recognized in later years that the conditions characterizing fully adjusted long-run equilibria are not usually met in the electricity area, with a reference to [3.13]: ”In recent years electric utilities on both sides of the Atlantic have been operating in a more-or-less continuous state of dis-equilibrium”.

The trouble with the models of Turvey is that indivisibilities are ruled out. If there are no indivisibilities, either regarding plant size or regarding construction and scrapping time, then there is no reason to make a distinction between short-run and long-run. Turvey naturally is aware of this, but is, like Boiteux preoccupied with the need of stable prices. The validity of that requirement is disputable. Not the validity of the requirement, but the way it was treated was contested by Kay [3.15]: “The most suitable method of deriving optimal prices is as the solution to some specified maximization problem in which the relevant constraints and assumptions are made explicit, not as the by-product of some arbitrary definition of marginal cost”. Kay illustrates this statement with the model described in Section 2.3.2. As stated there, one of the assumptions for the model is that demand is known beforehand, which is unrealistic. However, it was shown in Section 2.3.3 that the result with respect to *expected* revenue holds, even if uncertainty is taken into account. The assumption of varying prices may have been unrealistic in 1971, but is more acceptable today. Turvey stated in [3.5]:

- “… the solution (i.e. the design of dynamic tariffs, based on marginal cost and consumer willingness to pay) requires far more knowledge of elasticities and cross-elasticities than will ever be available.”
- “(…) the principles analysed in this chapter (Chapter 9 about more dynamic tariffs) will never be capable of precise application.”

The first statement is an argument for using market (SRMC-based) solutions: it will give consumers the opportunity to reveal these elasticities. Decreasing prices and technological development, reduce the validity of the second argument.

The dispute seems to be settled in favour of SRMC-based pricing. However, this does not solve the problem of providing sufficient peaking capacity as long as demand is inflexible. This was less of a problem in traditionally organized systems, but comes in the backdoor through the “revenue reconciliation” requirement. Andersson and Bohman treat this problem superficially only: “… it would be wise to (…) rely on pricing on SRMC with due consideration of budget constraints and other second best restrictions.” Brown and Johnsen also commented this in [3.7], where they conclude that “short-run operating costs will be recovered but capacity costs will not.” Their solution lies, interestingly in a futures market, where customers that recognize the probability of shortage can buy the right to be served. Through this mechanism they conclude a revenue balance can be secured.

3.1.3 Spot pricing of electricit y

In their book [3.16], Schweppe et al. develop a broad theoretical platform for the use of spot prices. Their motivation is:

- Freedom of choice (provide customers with options on cost and reliability)
- Economic efficiency (adjust electricity usage patterns to match marginal costs)
- Equity (reduce customer cross subsidies)

The main focus of the work is on a system with a regulated or government owned utility, but one chapter is concerned with a vision (in the mid-eighties) of total deregulation.

A very broad concept of spot price is presented. The spot price associated with the *k*th customer during the hour *t* is viewed as the sum of the following components:

$\rho_k(t) =$	$\gamma_F(t) +$	[Generation Marginal Fuel]
	$\gamma_M(t) +$	[Generation Marginal Maintenance]
	$\gamma_{QS}(t) +$	[Generation Quality of Supply]
	$\gamma_R(t) +$	[Generation Revenue Reconciliation]
	$\eta_{L,k}(t) +$	[Network Marginal Losses]
	$\eta_{QS,k}(t) +$	[Network Quality of Supply]
	$\eta_{R,k}(t)$	[Network Revenue Reconciliation]

Most of these components are readily applicable in a deregulated market. γ_F and γ_M result from generation bids, while the latter three components will be embedded into the network tariffs[3]. γ_{QS} and γ_R are more problematic. γ_{QS} can be handled in various ways, the most common being through reserve requirements that indirectly guarantee a satisfactory (though not necessarily optimal) level of security. Depending on the market organization, γ_{QS} may (like in the present UK solution) or may not (like in the Scandinavian solution) be embedded in the spot price. The treatment of γ_R is very problematic, and is closely connected with the whole peak-load problem. Indeed, Schweppe et al. state: “The world would be nicer if revenue reconciliation could be ignored”. They suggested three alternative solutions for the revenue reconciliation problem:

- Modifying the spot price
- Surcharge or refund
- Revolving fund

[3] This is by no means trivial, but not the theme of this thesis.

The spot price can be modified in several ways to ensure revenue reconciliation *ex ante* or *ex post*. This approach distorts spot prices, and does not give the correct short-term price signals to the market. The same is true for the surcharge/refund solution, at least when it is directly linked to energy usage[4,5]. Schweppe et al. recommend the use of the revolving fund solution, which means that it is accepted that revenues in some years exceed and in other years are less than costs. The advantage of this method is, naturally, that it does not distort spot prices, and allows for short-term efficiency. It is not clear from the argumentation how long-term revenue reconciliation can be guaranteed. The problem of the method in the context of [3.16] was that regulatory authorities did not allow it at that time. In a deregulated environment, excess revenues should be of no concern. At least in a well-functioning market, they should attract investment and be reduced to a natural level over time. Revenue deficiency though, is problematic when prices are inflexible.

Many authors in the technical sphere, have followed in the footsteps of Schweppe et al. The principle desirability of spot pricing as a means to social efficiency is not challenged. However, many have pursued ideas towards practical realization, revenue reconciliation and security concerns. The latter topic is typically within the engineering sphere: while economists are primarily concerned with economic efficiency, engineers are more engaged by the question of how to price security, with minimum efficiency loss.

Ideally, perfect spot pricing would involve more or less continuous (e.g. every second) metering of all consumers. The amounts of information exchange required with this ideal scheme are so enormous, that this is hardly feasible. A next step hourly or half-hourly metering, which still involves considerable amounts of information exchange and costs. With respect to practical realization, Baldick et al. [3.17] have proposed a scheme that catches some of the benefits of hourly spot pricing, without the same need for information exchange. They consider tariffs that i) are set in advance of participation, ii) are updated periodically, e.g. weekly, iii) consist of time sequences of prices to follow anticipated demand behaviour and iv) are long-lived relative to power systems time constants. Thus, these tariffs are somewhere between traditional time-of-use tariffs and spot prices. Still, the scheme requires considerable metering efforts (for example weekly), which are not compatible with traditional ways of meter reading. The authors do not explicitly assess the security problem. Their contribution is that they recognize that it is not possible to know the consumers' utility functions. They estimate these as part of their cyclical tariff updating process. Assuming quadratic utility functions (i.e. linear marginal willingness-to-pay functions), they prove local convergence to welfare optimality under some conditions of which the validity is not obvious.

[4] Schweppe et al. do not discuss the possibility of a non-linear tariff, consisting of a fixed amount in addition to the spot price.

[5] A possibility that is both efficient and satisfies the revenue reconciliation constraint is Auman-Shapley pricing, e.g. [3.21]. It will not be discussed further here, because it is a concept for a regulated environment or the regulated parts (e.g. the transmission network) of an otherwise deregulated environment.

Interestingly in the present context, the authors conclude that “the need for peaking plant may be significantly reduced” by using their pricing scheme.

The problem of revenue reconciliation is the subject of a paper by Kim and Baughman [3.18]. They compare four different approaches for revenue reconciliation in a case study for the Korea Electric Power Corporation: i) the adder method, where a constant term (possibly negative) is added to the spot price, ii) the multiplier method, where all spot prices are multiplied with the same factor, iii) the reliability sensitive adder, where the adder depends on the Loss of Load Probability, similar to the England and Wales solution and iv) Ramsey pricing, where prices for consumer segments with the lowest absolute price elasticity are increased most to minimize efficiency losses. These authors do not contribute with new pricing schemes, but with a test of reconciliation schemes on actual data. Their conclusions confirm existing knowledge like large deviations from marginal costs with the adder and multiplier models, largest price volatility with the LOLP method and greatest social surplus with Ramsey pricing.

Pricing of system security is the subject of a paper by Kaye et al. [3.19]. They extend the theory of Schweppe et al. by developing a framework for calculating the socially optimal level of security, thus resolving the security/economy trade-off. This is principally the same idea as used in Chapter 4 of this thesis. Kaye et al. define “contingency offerings”, defined as the additional power each participant (generator or consumer) could supply in the event of a contingency by increasing output or reducing demand respectively. A central dispatch is performed with the objective to achieve the socially optimal solution. It is shown that there is a price that can be set for contingency offerings and a forecast for actual contingent usage that encourages each participant to make a socially optimal contingent offering. The proposed approach has the characteristics that i) decision making is decentralized, ii) supply- and demand-side are treated equally, iii) utilities, other producers and consumers cooperate in system operation, iv) security-economy trade-off is incorporated and v) information technology is used to assist power system operation.

Pérez-Arriaga and Meseguer cover many of the issues in one single comprehensive model [3.20]. The paper looks at the pricing problem from the global perspective of generation expansion and operation. The model includes two new features: i) the viewpoint of competitive generation markets, where generators’ decisions are based on nonregulated market prices and ii) consideration of the effect of real operation and planning constraints, including security of supply in the short and long term. A major contribution is that model considers three time ranges for decision making: capacity expansion planning, operation planning and dispatch. It is shown that social, generator and consumer benefit are all maximized if prices equal marginal costs, including the effects of constraints on short and long-term security. Resulting prices exist of three components: the system marginal energy cost, the system marginal reserves cost and the system marginal cost of security of supply in the long term. It is also shown that long and short-term marginal costs are equal if there is no active reliability constraint, which is in the line of though of Williamson, cf. Section 3.1.1. It means that the optimal price may be lower or higher than short-term marginal cost + capacity (or reliability) cost, depending on presently available capacity. A practical implementation is suggested where the energy charge includes the first two components of the price, and an annual charge the third.

There is no doubt that “Spot pricing of electricity” by Schweppe et al. has had a significant impact on thinking about electricity pricing. There seems to be little disagreement about the book’s principal conclusions, but the more about practical ways to implement them.

3.2 The generation system and quality of supply

In the traditional view, demand has been regarded at price-inelastic in the short-term. The cost of load shedding has been estimated based on consumer surveys. The implicit result of this line of thought is that consumers prefer maximum obtainable reliability, and that this must be supplied by the generation system. Later in this chapter, this view will be challenged, but in this section the generation system as the supplier of reliability will be discussed first.

3.2.1 Power system security

In Chapter 2, the concept of power system reliability was discussed. It was pointed out that the very broad concept of reliability is subdivided in system adequacy and security. System adequacy relates to the existence of sufficient facilities, and is associated with static conditions, while system security has to do with the system’s ability to respond to disturbances. Most of the research within the field concerns system adequacy, and some of these results are shortly referred in Chapter 2.

The concepts of adequacy and security are related: the level of adequacy constrains the obtainable level of security. If system adequacy is insufficient, it may be impossible to obtain the desired level of security, because the necessary generation capacity to obtain this level of security is not available. In a nutshell, this is the quintessence of the peaking capacity problem in restructured electricity markets.

The conventional way to maintain security at the generation level is through the use of fixed reserve requirements, discussed in the next section. A more dynamic approach was adopted by the Pennsylvania - New Jersey - Maryland (PJM) Interconnection [3.27]. In this approach, the principles of adequacy assessment are used to calculate spinning reserve levels that result in a constant level of security, independent of load level or load forecast uncertainty. From a theoretical point of view, this is an attractive approach, under the assumption that the value of lost load is equal in all hours. (If this were not the case, it would be easy to correct the method.)

The PJM method is an interesting basis to assess situations where available capacity is insufficient to apply agreed-upon reserve levels. By analyzing all outage situations (with a probability over some threshold), and their resulting loss of load, it is in principle possible to balance outage costs against the immediate cost of load shedding to provide sufficient reserves.

3.2.2 Reserve requirements

In a traditional planning environment, one way to assess system adequacy was the use of a fixed reserve requirement. For several reasons, this is an unsatisfactory way to assess the problem of system adequacy [3.25], and more advanced methods, using probabilistic criteria are now mostly applied.

In principle, some of the objections against the use of fixed reserve requirements for adequacy assessment, are also valid when security is addressed. However, in the latter case, the use of fixed reserve requirements is dominating. (Some notable exceptions are [3.27] and [3.28].) Some reasons for the difference in approach between adequacy and security assessment are:

- Comprehensive calculations are necessary for probabilistic reliability assessment. This is feasible in the off-line environment of adequacy calculations, but problematic in a real-time situation.
- The cost of keeping excessive reserves under low-load conditions is not necessarily high. So instead of varying reserve requirements depending on the load situation, the peak load requirements are used continually.
- At present, many power systems are an integrated part of very large systems, with a huge inertia, where during low load conditions reserve may be available almost for free.

Reserve requirements are normally divided in several categories. In the European tradition, a common classification is between primary, secondary and tertiary reserves (cf. [3.29] Appendix 3). In the US tradition, it is common to use the terms regulation and spinning, supplemental and backup reserves. Though there are numerous differences in philosophy and implementation between various systems, there are some general characteristics:

- Primary/spinning reserves are frequency controlled, and used to cover loss of the largest unit. They will also take care of normal load fluctuations and minor load forecast deviations.
- The role of secondary/supplemental reserves is to move the frequency set point back to its basis, after a deviation has occurred through the use of primary reserves. The indication for the use of secondary reserves may be an Area Control Error or time deviation. Control is by Automatic Generation Control in many systems, but can also be manual. Secondary reserves may be spinning or non-spinning, depending on the system and plant characteristics.
- Tertiary reserves are used to replace secondary reserves when the usage of the latter has made this necessary. The rules concerning tertiary reserves seem to be less stringent than for primary and secondary reserves, cf. [3.29].

Given some classification of reserves (a finer subdivision than this may well be possible), it is important to decide the optimal level of each category. In the US, regulation is “the use of generation (…) to maintain minute-to-minute generation/load balance within the control area to meet NERC control-performance standards” [3.30]. The size of the largest unit is the main ingredient in the requirement for primary reserve. The amount of secondary reserves is less well defined. UCTE uses recommended values equal to the numeral value of $3 \cdot \sqrt{P_{L,MAX}}$, while NORDEL leaves this to each country.

Thus, the generation system provides security to avoid loss of load by maintaining reserve levels at least as high as the agreed upon requirements. From a theoretical point of view, these requirements are not very satisfactory, but they are workable in practice, and probably the potential savings by using better methods are small during normal operating situations. This changes in situations where there is insufficient capacity to provide for demand and reserves. In these cases the difficult question emerges if reserve requirements are hard constraints, resulting in load shedding to provide reserves, or soft constraints, resulting in lower-than-planned system security. To make a reasonable choice between these alternatives, it is necessary to take a closer look at the demand side. This is done in Sections 3.3 and 3.4.

3.2.3 The public good view

In a restructured market, it is not obvious how a necessary level of system security can be maintained, and how the "producers of system security" should be compensated. In this context it is interesting to look at the theory for public goods, cf. [3.22], Chapter 25. There are two basic properties that characterize public goods:

- Nonexclusivity
- Nonrivalry

A good is nonexclusive if it is impossible, or very costly, to exclude individuals from benefiting from the good. Examples are clean air or street lighting. A good is nonrival if consumption of additional units of the good involves zero social marginal costs of production. Examples are use of television channels or bridges (at least below their capacity ...). Nonexclusivity is the dominating characteristic for public goods. In [3.22], a good is defined as a public good if "… once produced, no one can be excluded from benefiting from its availability. Public goods usually also will be nonrival, but that need not always be the case."

In the conventional view, system security is nonexclusive: if the system has a certain level of security, all consumers benefit from it. The nonrivalry property is irrelevant from the consumer point of view: as long as security is assumed non-differentiable, no consumer can choose to increase or decrease his consumption of it. From the generation point of view, it may be argued that the nonrival property does not hold: e.g. the installation of a unit that is larger than any of the existing units incurs increased costs to maintain the same security. However, because of the nonexclusivity property, system security is a public good.

The well-known problem with public goods is that their production will be less than the social optimum if left to the market, which is a consequence of their nonexclusivity. In such cases, the good is often provided by public bodies, which cover their expenses through general taxes or more dedicated charges. In the case of system security, this is obtained by letting the System Operator procure the necessary services from generators, and divide the costs over all users of the system, for example as a part of the transmission tariff. A difficult challenge in this case is to find the optimal level of security, because there is no price mechanism that could result in equilibrium. This becomes especially cumbersome when lack of generation capacity makes it impossible to provide the security level that the system operator requires.

The public good view is another argument for the hypothesis that an energy-only electricity spot market *without additional mechanisms* will not enough generation capacity in the long run.

3.3 Consumer willingness to pay for electricity

In a basic spot market setting, given enough competition, prices will settle down at marginal cost. The discussion in Section 3.1.1 should make it clear that there has been controversy about the correct marginal cost to be used in a traditionally regulated environment. While some argue this should be LRMC, others advocate SRMC, while it is also claimed that these are actually equal.

In a restructured system, the controversy seems to have been settled in favour of SRMC, at least as far as generation, the theme of this thesis, is concerned. This is surely the case in the Nordic and the Californian markets, where no other generation related price elements than the generator's bids are incorporated[6]. However, based on considerations of long-term system security, other implementations have added several forms of non-market based price elements. These may be seen as movements in the direction of "ƅ+β" instead of "ƅ" in the terminology of Section 3.1.1.

In this section, focus will be on consumer willingness to pay for electricity, and its implications for a short-term efficient spot market solution with the proper level of system security.

Consumer utility is based on the concept of "preference"[7]. The preference relation is usually assumed to have the three basic properties of completeness, transitivity and continuity. Under these assumptions, it is possible to show formally that people are able to rank "commodity bundles" in a desirability order, in such a way that more desirable bundles offer more utility than less desirable bundles. It is also possible to attach numbers to this utility, but an important point is that these numbers are not unique. They are defined only up to an order-preserving (monotonic) transformation. This introduces difficulties if one would attempt to use this concept to find an actual demand function based on utility, because any strictly increasing transformation would yield an equally valid utility function, but a different demand function. So following other presentations in the field, the starting point for the analyses will be a "willingness to pay" or inverse demand function, giving price as a function of volume.

[6] The Californian market has advanced markets for ancillary services, but these are still short-term markets, in principle based on short-run costs for production of energy and these services.

[7] A basic introduction, satisfactory for the current context is given in [3.22]. A broader and more theoretical discussion is found in [3.23].

The basis assumptions in this section are:

- Consumers receive utility from using electricity
- Their willingness to pay is non-increasing, i.e.

  $\partial q(d)/\partial d \leq 0$

  where q is the willingness to pay and d is the power consumed
- The time horizon is one hour[8], which naturally excludes investments
- System security is taken care of by fixed reserve requirement
- The objective is to maximize short-term welfare, while satisfying the constraints

The problem is to find an optimal solution in cases where demand is capacity constrained. Mathematically, the problem may be stated as:

$$\underset{p_0}{\text{MAX}} \quad W = \int_0^{p_0} [q(x) - c(x)] \; dx \qquad \textbf{(3-1)}$$

subject to:

$$p_0 = d_0 \qquad \textbf{(3-2)}$$

$$P_{MAX} - p \geq R \qquad \textbf{(3-3)}$$

$$p_0, d_0 \geq 0 \qquad \textbf{(3-4)}$$

where

W	- social welfare
P_{MAX}	- maximum production
R	- reserve requirement
p_0	- generation
d_0	- demand
$q(.)$	- consumer willingness to pay
$c(.)$	- marginal generation cost

This in principle simple, one-period one-generator problem serves to demonstrate various approaches. The single generator may be thought of as representing a great number of generators, sorted after increasing costs. The resulting increasing cost function *c(.)* will in this case include the effects of unit commitment, given the situation at the start of the hour considered. Reserves are represented very basically in this principle exposition.

[8] Some systems operate with shorter commitment and/or dispatch periods. Here “hour” will be used, but there is no principal difference for other period lengths.

3.3.1 Elastic demand

With elastic demand $q_1(d)$, the problem becomes simple, Figure 3-3. The solution is given by $p_0 = d_0 = PMAX - R$, with the price $q_1(d_0) \geq c(p_0)$. The difference $q_1(d_0) - c(p_0)$ is equal to the dual value of the constraint **(3-3)**. It represents a scarcity rent to the producers, which over time should motivate to investment in new capacity.

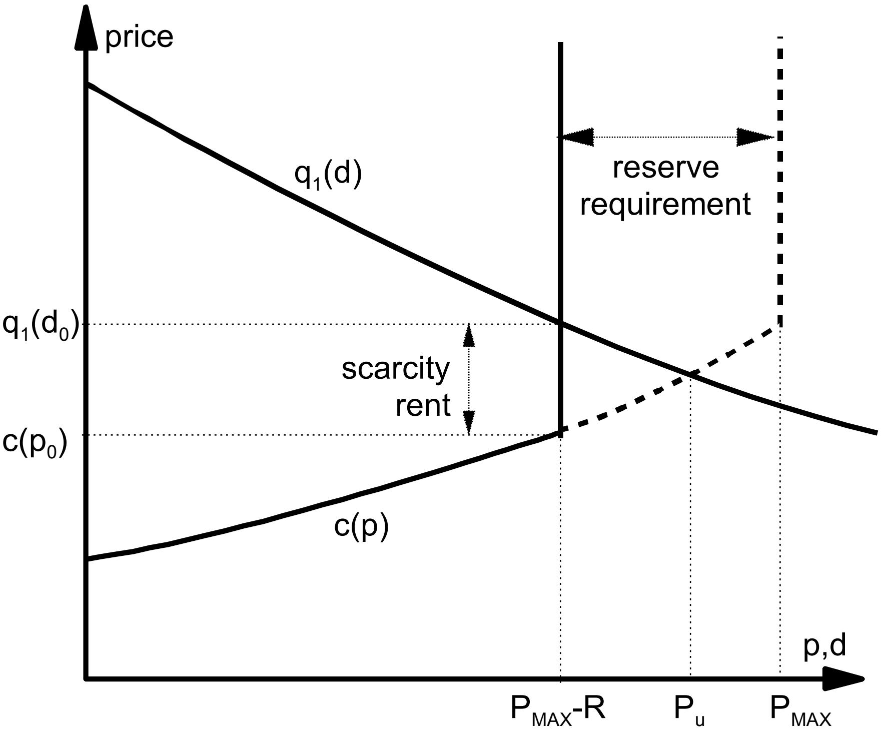

Figure 3-3: Capacity constraint with elastic demand

This is the standard way to solve the problem in economic theory. However, the assumptions are unrealistic: demand is not currently elastic in the very short run. Consequently, due to technical and organizational limitations, demand will not be reduced in the very short run, however high prices become. As a result if this, the system operator will have the choice between reducing system security below agreed-upon standards, or resort to load shedding (rationing).

3.3.2 Inelastic demand

Most consumers in restructured systems still buy their electricity from a supplier, who in turn buys from a generator, possibly through one or more intermediaries[9]. Consumers normally have a form for tariff, where the short-term price they pay is independent of the short run

[9] In several cases (e.g. the Nordic and English), one or more pre-defined load-profiles are used for settlement of consumers buying electricity from a supplier that has no direct access to their meter. In Norway, some suppliers even offer contracts based on spot prices plus a markup. This creates a somewhat odd situation for the consumer: in theory he pays the spot price, but once he has entered into the contract, his actual hourly consumption hardly influences his costs at all. Consequently, he has no motivation to reduce consumption when prices are high.

marginal cost (including scarcity rents) of the system. In this way, their consumption becomes de facto inelastic compared with short run marginal cost.

Inelastic demand is a somewhat troublesome notion when used in a welfare maximization context. When demand is completely inelastic, the demand curve is vertical and maximization of the sum of consumer and producer surplus makes no sense. It seems reasonable that there is some limit to the inelasticity. For example, when prices become extremely high, suppliers may do everything in their power to appeal to consumers to reduce their loads. However this is accomplished, it is reasonable to use an upper price limit, called λ_{sh}[10], where the index *sh* stands for "shedding". The value could be set to the Value Of Lost Load, *VOLL*, roughly 4 $/kWh in the UK case. The resulting inverse demand function *q2(d)* is shown as a solid and striped-dotted line in Figure 3-4.

Another way to solve this problem is by observing changes in social welfare, compared with some base case, which makes it unnecessary to adopt the arbitrary limit λ_{sh}. This should give the same result, but the former solution is more convenient in the current context.

Under these assumptions, it is obvious from Figure 3-4 that the solution of **(3-1)** is $p_0 = d_0 = PMAX\text{-}R$, as before, while the price equals $q_2(d_0) = \lambda_{sh}$.

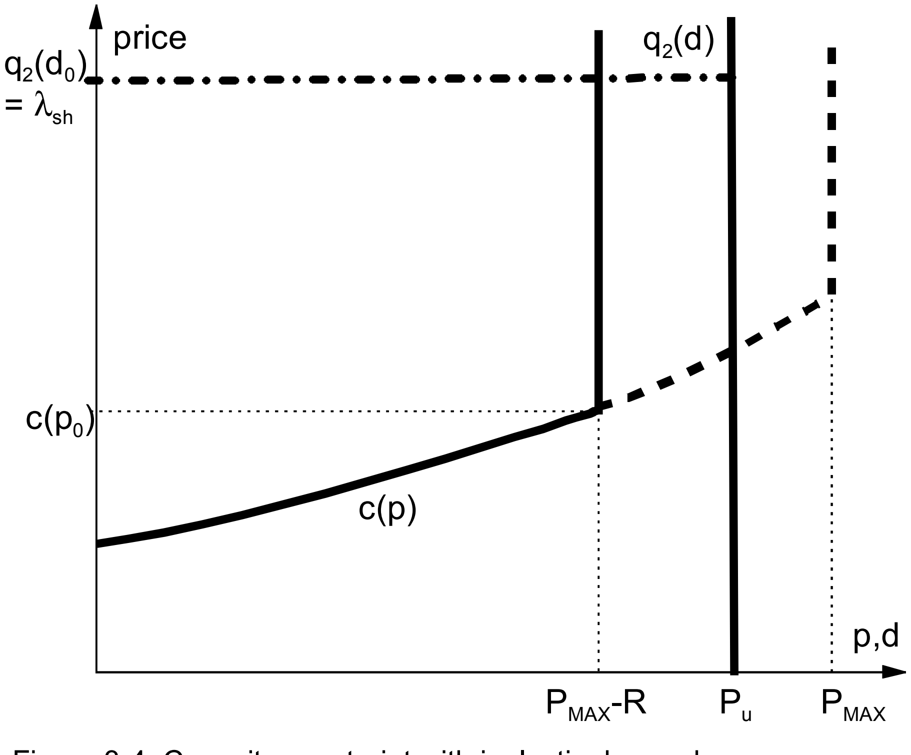

Figure 3-4: Capacity constraint with inelastic demand

Because all consumers are assumed to experience the same inconvenience of involuntary rationing, random rationing will be the most rational way to obtain a market equilibrium. Assuming the price is set administratively to λ_{sh}, consumer surplus is zero, while producer surplus is very high. A market price at this level may create political problems of acceptance,

[10] When writing about power and economics there is a symbol problem: while p always will be associated with price by an economist, the electrical engineer will automatically think of power. In this thesis, the symbol p is used for power, while λ is used for prices.

even if it has a short duration. A household buying 5 kW on the spot market for 10 hours with this price, will get a bill of 200 $. A much worse situation may be faced by companies unable to reduce their consumption or unaware of prices. On the other hand, the income created, 40 $/kW, may only be just enough to invest in gas turbines. Keeping prices lower for political reasons will deter investments.

A milder version of inelasticity is obtained by dividing consumption in groups with assumed similar values of lost load. In this case the groups with lowest values can be rationed first. However, this presumes that it is possible to switch off consumers according to these criteria.

3.3.3 Seemingly inelastic demand

The situations in the previous two sections are extreme. It is plausible to assume that there is some short-term price elasticity, but that technological and organizational obstacles obscure this. We assume that there is some short-term price elasticity, but that technological and organizational obstacles obscure this. The situation is shown in Figure 3-5.

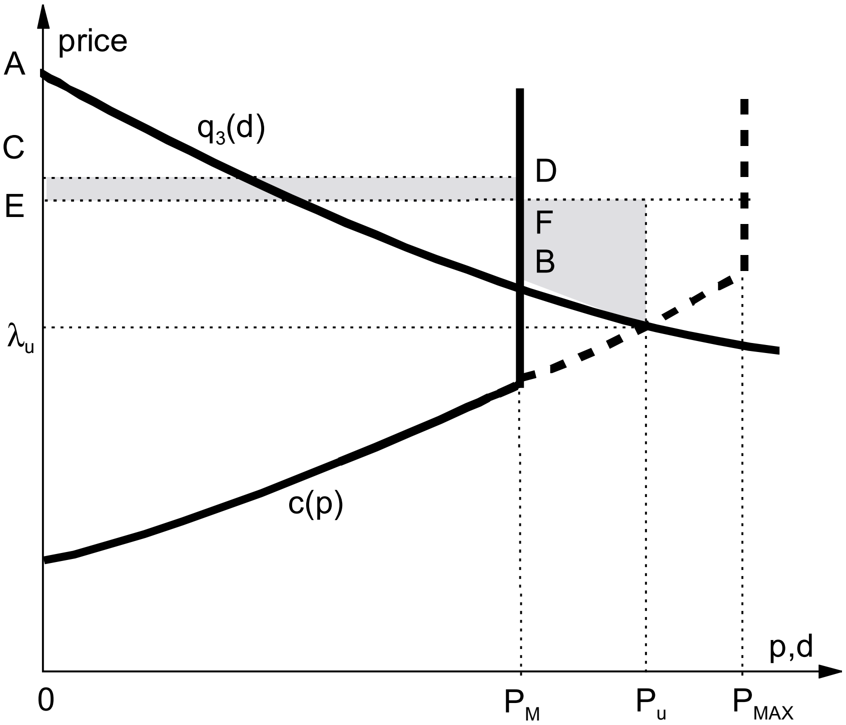

Figure 3-5: Capacity constraint with seemingly inelastic demand

When no capacity constraints are active, the market will settle at an unconstrained price λ_u and satisfy a demand of P_u. If the capacity constraint is active, demand must be reduced to $P_M = P_{MAX} - R$. Because the real elasticity is invisible for the system operator, random rationing is the only possible choice. A better solution would be to ration consumers according to their willingness to pay. The welfare gain of this, net of the cost of implementation, is given by the difference in total revenue plus consumer, because total generation is identical (P_M) in both cases. In Figure 3-5, total revenue plus consumer surplus with perfect rationing is represented by the area 0ABP$_M$, which is equal to 0CDP$_M$, where CD is the average value of *$q_3(d)$* over 0P$_M$. With random rationing, this is equal to 0EFP$_M$, where EF represents the average value of q$_3$(d) over 0P$_u$. Consequently, the gain is given by ECDF and can be expressed by:

$$W_{gain} = \int_0^{P_M} q_3(x)dx - \frac{P_M}{P_u}\int_0^{P_u} q_3(x)dx \qquad \textbf{(3-5)}$$

where the first term represents the area 0CDP$_M$, and the second the area 0EFP$_M$. Alternatively, the gain can be viewed as the difference between the lost surplus with random rationing and perfect rationing respectively. The two equally large shaded, areas in Figure 3-5 represent each of these views. It is easy to see that if *$q_3(x)$* is constant, CD and EF coincide, and W_{gain} is zero. In this case random rationing is as good as any other way, and cheapest to implement.

To further study the properties of this model, the piecewise linear demand function *$q_4(d)$* shown in Figure 3-6 is assumed, where the shaded areas correspond to those in Figure 3-5

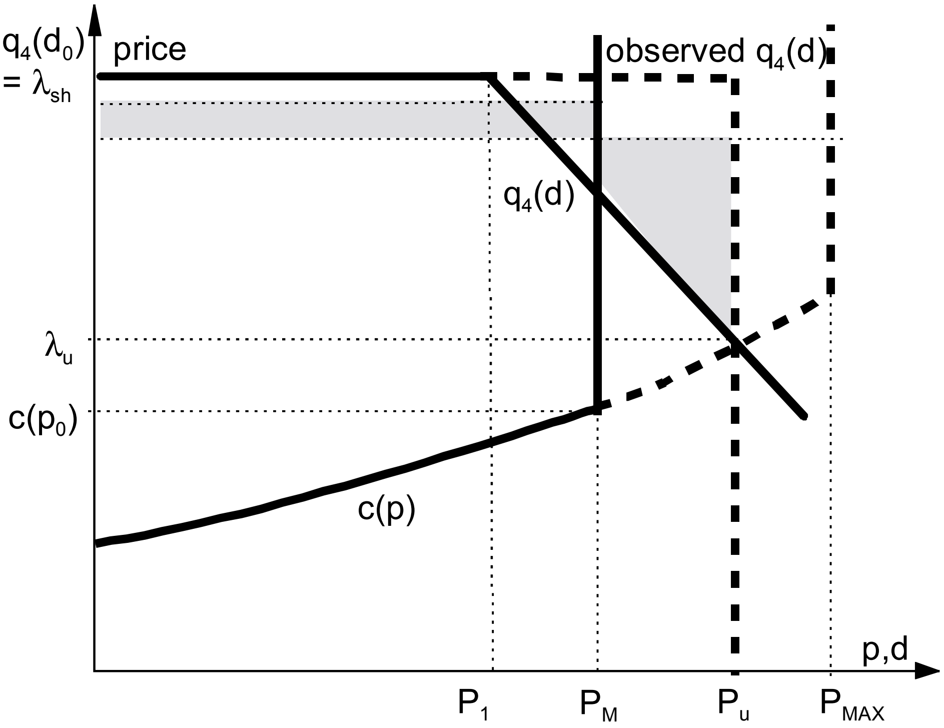

Figure 3-6: Piecewise linear demand function

Up to *P_1* demand is inelastic, which is represented by perfect elasticity at λ_{sh} = *VOLL*, like in 3.3.2. For higher values of *p*, a linear demand function is assumed. In this case, the expression in **(3-5)** becomes:

$$
\begin{aligned}
&W_{gain} = W_1 - W_2 - W_3 \\
&\text{where} \\
&W_1 \quad = \lambda_{sh} P_M \\
&W_2 \quad = \frac{\lambda_{sh} - \lambda_u}{2(P_u - P_1)} (P_1 - P_M)^2 \qquad \textit{if} \quad P_1 \le P_M \\
&\qquad\;\; = 0 \qquad\qquad\qquad\qquad\qquad\; \textit{if} \quad P_1 > P_M \\
&W_3 \quad = \frac{P_M}{2P_u} \left[\lambda_{sh} \left(P_u + P_1 \right) + \lambda_u \left(P_u - P_1 \right) \right]
\end{aligned}
\qquad \textbf{(3-6)}
$$

W_1 - W_2 represent the first term in **(3-5)**, while W_3 represents the last. For $P_1 = P_u$, $W_2 = 0$, $W_1 = -W_3$, and W_{gain} becomes zero: this shows again that if demand is really inelastic, random rationing is the best and only realistic alternative.

Given that P_1 is uncertain, it is of interest to find out how the gain of perfect compared with random rationing depends on P_1. The value of P_1 for which the maximum occurs is found by derivation, after some manipulation resulting in:

$$P_1 = P_u \pm \sqrt{P_u (P_u - P_M)} \qquad \textbf{(3-7)}$$

The maximum[11] welfare loss occurs for the minus sign (because otherwise a value greater than P_M would occur). Normally P_u - P_1 is small compared with P_u, which means that the maximum welfare gain will occur for relatively high values of P_1. Probably, P_1 is rather high in reality, which means that random rationing (if capacity shortage occurs) yields relatively high welfare losses.

To get an idea of the size of potential gains, some tentative numbers for the Norwegian situation may be used:

P_u	= 24000 MW
P_M	= 23000 MW
λ_{sh}	= 22000 NOK/MWh, cf. [3.24]
λ_u	= 250 NOK/MWH

With P_1 = 20000 MW, hourly welfare gain from using the latent underlying price elasticity instead of using random rationing would be 17.2 million NOK per hour. If such a situation would occur 50 hours each year, annual welfare losses would amount to 860 million NOK. Figure 3-7 shows how potential welfare gains depend on P_1 in this case for some values of the unconstrained demand P_u. The figure shows clearly that, if only a minor part of demand is short-term elastic, large welfare gains can be made from exploiting this elasticity, compared with random rationing.

[11] It is straightforward to show that the second order conditions for a maximum are satisfied.

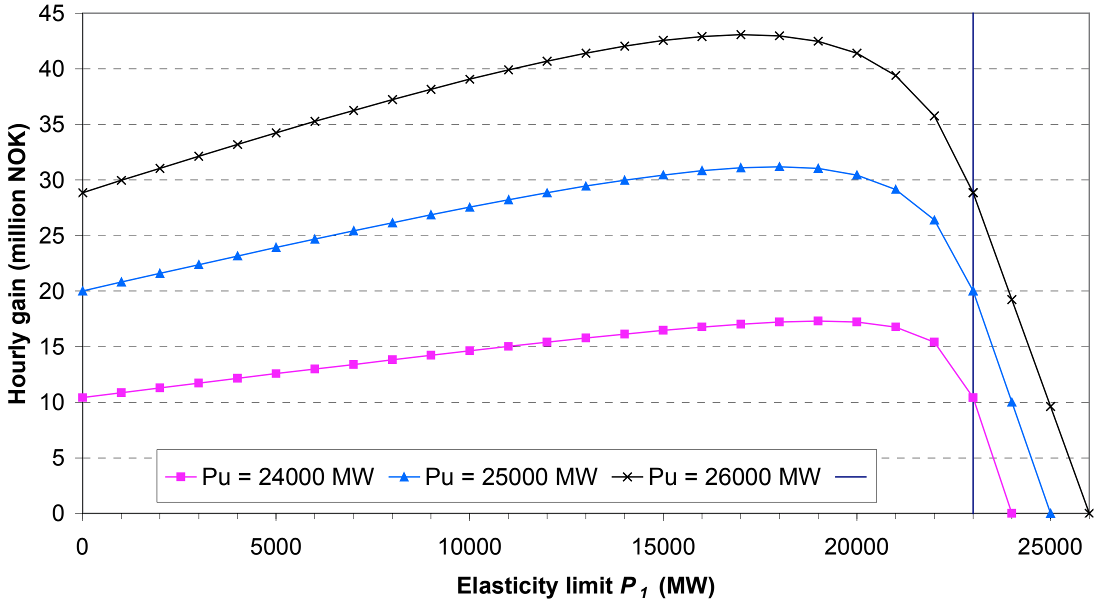

Figure 3-7: Hourly welfare gains as a function of the value below which short-term demand is inelastic, P_{MAX} = 23000 MW

Practical implementation poses some challenging questions with respect to the division of social welfare between consumers and producers. If demand elasticity is exploited in such a way that a number of consumers actually bid in the spot or regulation market, normal demand elasticity will be obtained, and the situation described in 3.3.1 emerges. If, however, this is accomplished through agreements between the System Operator and elected consumers, based on fixed annual payments, rationing may be available at a very low *marginal* cost. Prices will stay relatively low (at least compared with shedding cost), and producers will not receive any scarcity rent[12]. In other words, the method of implementation has significant impact on the distribution of social surplus between consumers and producers, and consequently on investment behaviour.

3.3.4 Demand used as reserve capacity

So far, it has been assumed that a fixed amount of reserves has to be provided by the generators. This means that available generation capacity is equal to $P_M - R$. When unconstrained demand, with given pricing structures, exceeds this value, it has to be constrained in some way. If demand could be used to provide some of the reserves, total welfare losses might be reduced, and prices could be lower.

With inelastic demand, as in 3.3.2, consumer's short-term willingness to pay is equal to *VOLL*, and welfare loss is equal to *VOLL* multiplied with disconnected load. So in case of a capacity shortage of $P_u - P_M$, short-term welfare loss is simply $VOLL \cdot (P_u - P_M)$. If however a

[12] Similar situations have happened in the Norwegian system in 1999, where Statnett chose to disconnect a number of large boilers, with the result that spot prices stayed low. Some producers criticized this.

share α of demand can be used as reserve, and the probability of activation[13] of this reserve is γ, welfare loss could be reduced to *$(1 - (1-\gamma)\cdot\alpha) \cdot VOLL \cdot (P_u - P_M)$*. How great this reduction would be, naturally depends on the values of α and γ. The value of α depends on technological feasibility and consumer's preferences for certain disconnection and disconnection with probability γ respectively. The value of γ depends on system characteristics and the position of demand side reserves in the "reserve merit list".

If part of demand is elastic, like in 3.3.3, welfare gain by using demand as reserve capacity is given by:

$$W_{rgain} = \begin{array}{l}(1-\gamma)\displaystyle\int_{P_M}^{\min(P_M+\alpha P_u, P_u)}[q(x)-c(x)]dx \\ +f_R(\alpha(P_u - P_M))\end{array} \qquad \textbf{(3-8)}$$

Here the last term *$f_R(.)$* represents the generator's cost of providing the same reserves. The symbol *W_{rgain}* is used to distinguish from the gain calculated in 3.3.3.

When *$q(d)$* is specialized as in Figure 3-6 and a constant generation cost equal to λ_u is assumed, results may be calculated in a similar way as in **(3-6)**. The formulas become somewhat messy, and depend on the position of *P_1* compared with *P_M* and *P_u*.

With the numbers from the previous section, and furthermore assuming $\alpha = 0.5$ and $\gamma = 0.1$, welfare gain from using demand as reserves is 2.7 million NOK per hour.

Figure 3-8 shows how these gains depend on demand elasticity, as defined in **(3-6)**. The picture is very different from Figure 3-7. For reserves, the largest gains will occur with low demand elasticity. When using demand elasticity for actually switching off demand, the largest gain (compared with random rationing) will be obtained for relatively high elasticity, and stay at a high level even for very large demand elasticity (when $P_1 = 0$).

[13] In the sense of activated volume divided by available volume

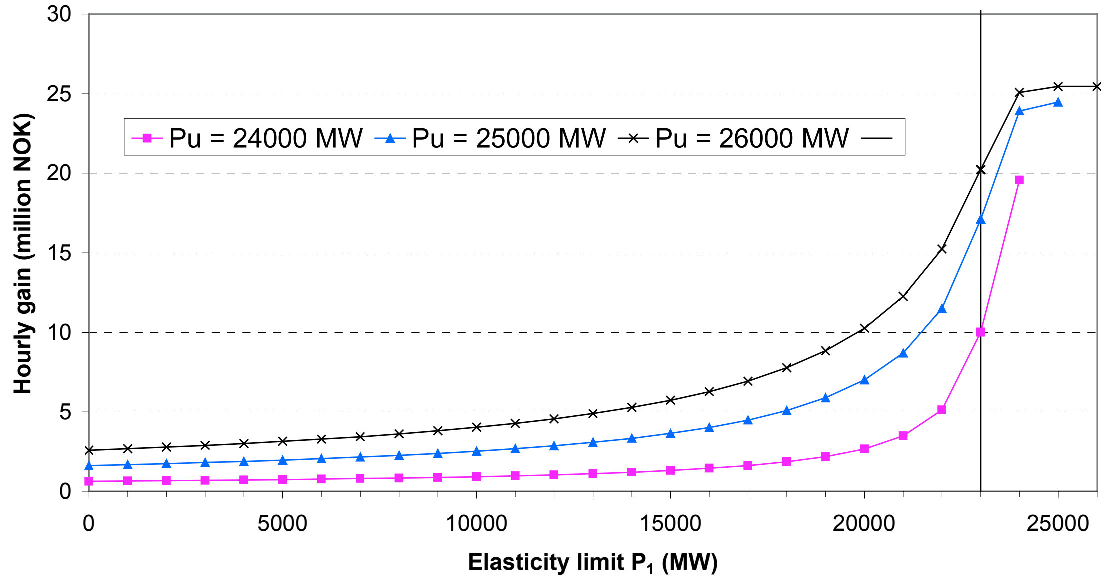

Figure 3-8: Hourly welfare gain of using demand for reserves as a function of the value below which short-term demand is inelastic; P_{MAX}=23000 MW, α=0.05, γ=0.1

Finally, shows the dependency of the welfare gain on the share α of demand being willing to serve as reserve. The interesting point is that there is a considerable gain, even if only a small part of demand is willing to stand as reserve.

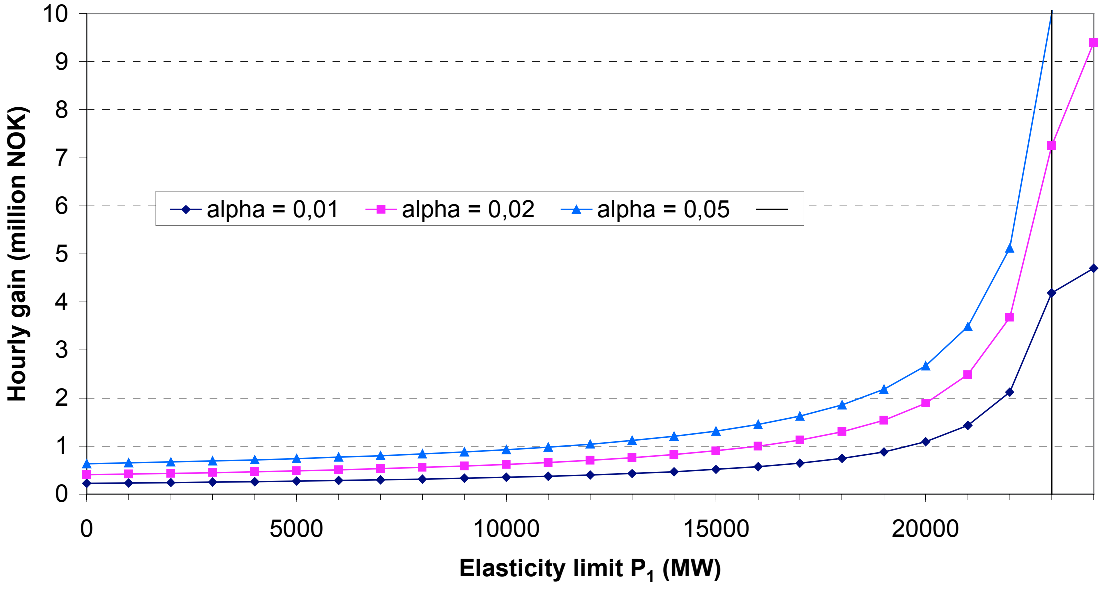

Figure 3-9: Hourly welfare gain of using demand for reserves as a function of the value below which short-term demand is inelastic; P_{MAX}=23000 MW, P_u=24000 MW, γ=0.1.

3.4 Willingness to pay for electricity and quality of supply

A basic assumption in Section 3.3 was that there exists some externally imposed reserve level that must be satisfied during system operation. Although consumers could supply some of the reserves in 3.3.4, the basic assumption of a given necessary reserve level was not abandoned.

The main role of reserves is to provide security against loss of load. Without reserves, minor deviations from demand forecasts or plant failures would result in partial or total loss of load, with huge welfare losses. Reserve levels have been determined historically, based on various criteria like Loss Of Load Probability (LOLP).

Theoretically, the optimal level of reserves is such that the marginal cost of Expected Energy Not Served (EENS) equals the marginal cost of increasing system security. The difficult part of this criterion is calculation of the operational value of EENS, which requires evaluation of huge numbers of fault situations. A model that takes this approach is described in Chapter 4. Here the view is taken that consumers are not interested in reserves per se. It is assumed that consumers are interested in two "products":

- electricity
- quality of supply

Several authors have reasoned along this line of thought. Carlton [3.10] develops a model for a good with two characteristics, its price and the probability that it can be purchased. Customers have preferences not only for the good itself, but also for the probability of obtaining it. In the market equilibrium derived by Carlton, there is a positive probability of not obtaining the good. The price of the good exceeds the cost of production, because the revenues must compensate for the cost of unused capacity.

Meyer [3.11] takes the view that quality of the supply, in the sense of the probability of being served is an important property of the product when supply is not 100 % guaranteed, and that this property affects utility.

Tschirhart [3.31] also recognizes the relation between demand and reliability, by defining a demand function *X(p,ρ)*, where *p* is the price of electricity and *ρ* is the obtained reliability. A simple example of the form

$$X(p,\rho)=\frac{(A-p)\rho}{\rho_0}$$

is used, where *A* and ρ_0 are constants.

In a more recent publication, Ravid [3.32] develops a model where consumers receive welfare from the consumption of electricity and from the consumption of reliability. Ravid uses a simple additive model for consumer utility:

$$U=\overline{U}(p)+k\alpha(k)$$

where $\overline{U}(p)$ is utility from consuming electricity, *k* the price of reliability and *a(k)* reliability. The price for reliability appears to be such that the marginal "cost" associated with increased reliability in the market for electricity (i.e. changes in electricity prices or capacity level) will be equal to the marginal associated welfare gain in the market for reliability.

This section will deal with an alternative quantification of “quality of supply”. A demand model for the two products will be proposed, and its properties will be discussed.

3.4.1 Quantification of quality of supply

The concept of “quality of supply[14]” is assumed limited to loss of load. This implies, that properties like voltage and frequency are not taken into account: either they are assumed within their defined quality levels, or consumers are assumed not to bother about these aspects.

It is assumed that consumers are interested in (or receive utility from) electricity and quality of supply. The latter may be quantified in several ways, e.g.:

• Loss Of Load Probability[15]	*LOLP*
• Expected Energy Not Served	*EENS*
• Energy Demand minus EENS	$d_w - EENS$
• Energy Index of Reliability	$EIR = (d_w - EENS)/d_w = 1\text{-}LOLP$

LOLP and *EENS* are quantities that consumers want as little as possible of, i.e. they are economic bads. It is more convenient to use economic goods, i.e. goods that consumers do want. Demand minus *EENS* is an economic good, but its value depends on absolute electricity consumption. This means that a consumer with high demand experiences a higher quality of supply than a consumer with low demand, given the same system *EENS*. By dividing this value by demand, the Energy Index of Reliability *EIR* is obtained, which looks suitable to represent quality of supply. It has one potential drawback with regard to further application of demand theory: its value can never exceed 1. This means that demand for quality of supply in isolation is not locally insatiable: there is some finite value, 1.0, above which it does not make sense to demand more. Because of this, one might have doubts if Walras’ law $px \le Y$ (price multiplied with demand is less or equal to income, where x=(electricity, quality of supply)[16]) holds with equality. However, consumers may still receive more utility from receiving more electricity, which means that in the vicinity of any point (p,x), there will still exist another point (p',x') that consumers will prefer, even if quality is at its limit. If the difference between electricity demand and *EENS* is divided by *EENS* instead of electricity consumption, the resulting quantity also increases with decreasing *EENS*, and has no finite limit. But this advantage does not seem to outweigh the slightly more messy mathematics.

As a conclusion to this discussion, quality of supply is defined as:

[14] Further on “quality” will be used when this is unambiguous.

[15] In this section, LOLP is used as probability index and not as a frequency, cf. Section 2.2.1

[16] Cf. [3.23], page 41.

$$qs = EIR = \frac{d_w - EENS}{d_w} \qquad \textbf{(3-9)}$$

where d_w is demand for electricity with 100 % reliability, *EENS* is expected energy not served.

3.4.2 A utility function for electricity and quality of supply

In the previous section, it has been argued that consumers receive utility from electricity and quality of supply. The question now is what a utility function should look like. It is obvious that the "real" utility function is hard to find. Its assessment should be based on hypothetical comparisons for example between receiving d_w kWh with 95 % reliability or 0.8 · d_w with 98 % reliability. No research has been done to this respect to the author's knowledge.

Presently, ambitions must be limited to finding functional relations that yield plausible short run willingness to pay functions for electricity and quality, and that can shed light on principle aspects. A conventional Cobb-Douglas function is used:

$$B(d_w, d_{qs}) = d_w^{\;\beta 1} \cdot d_{qs}^{\;\beta 2} \qquad \textbf{(3-10)}$$

or:

$$\frac{B(d_w, d_{qs})^{1/\beta_1}}{w_1} = \frac{d_w}{w_1} \cdot d_{qs}^{\;\beta 2/\beta 1} \qquad \textbf{(3-11)}$$

where *B(.)* is consumer utility, d_w and d_{qs} are consumed levels of electricity and quality, β_1, β_2 are coefficients and w_1 is some arbitrary reference value

Figure 3-10 shows some iso-utility curves **(3-11)** for β_1/β_2 = 1.0 (reasonable values for β_1 and β_2 will be discussed in the next section). The value for w_1 is chosen such that a quality of 1.0 corresponds to an energy consumption of w_1 for curve number 2. A consumer with these preferences would receive the same utility from an energy consumption of w_1 and a quality of 1.0 as from an energy consumption of 1.25w$_1$ and a quality of 0.8. The figure also shows the "budget constraint" which requires some discussion in this context. Many surveys of willingness to pay to avoid load-shedding have been conducted internationally, cf. [3.24]. The objective of these surveys is to find the Value Of Lost Load, *VOLL*. It is reasonable to assume that this value indicates consumers' willingness to pay for 100 % reliable supply of electricity in the short run, provided quality does not cost anything. (Because of the fact that electricity by most consumers in the industrialized world is conceived as close to 100 % reliable). Thus the budget may be set equal to this value: it is the maximum part of his income the consumer is prepared to spend on electricity. With this assumption, optimal utility is obtained for d_w/w_1=0.5 and d_{qs}=0.7 in

Figure 3-10.

The slope of the budget constraint corresponds to the price ratio of electricity and quality. If the price of the latter is 0, the optimal solution is to obtain a quality of 1.0 and energy equal to w_1, touching curve 4. The quantity variations have been exaggerated in this figure to show the principal situation.

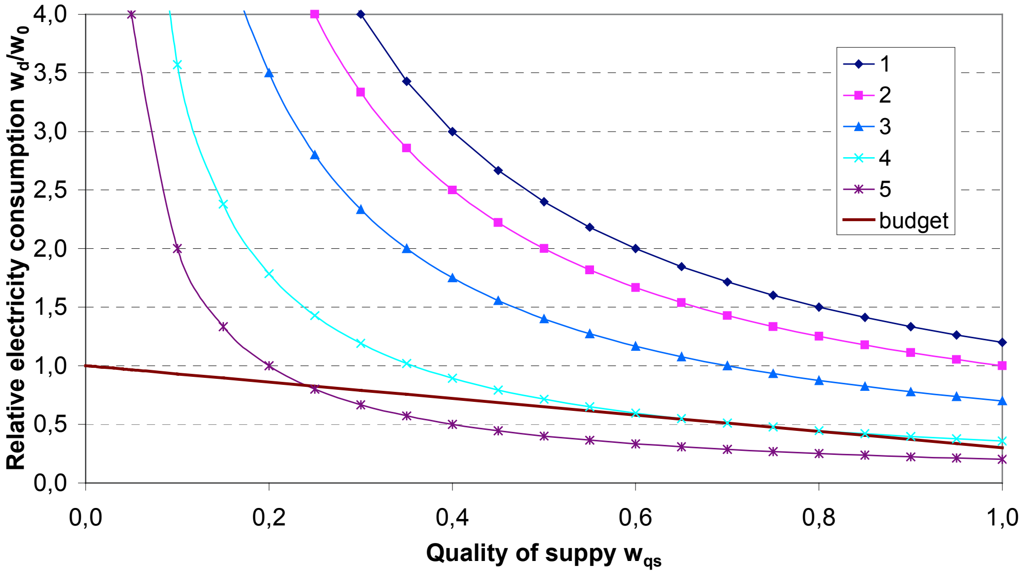

Figure 3-10: Iso-utility curves for electricity and quality of supply

If prices for electricity and quality are given, i.e. if the consumer is a price taker in a competitive market, the consumer problem is to optimize his utility **(3-10)** or **(3-11)**, subject to the constraints:

$$d_w \cdot \lambda_w + d_{qs} \cdot \lambda_{qs} = I \qquad \textbf{(3-12)}$$

$$d_{qs} \leq \overline{qs} \qquad \textbf{(3-13)}$$

$$d_w \leq \overline{w} \qquad \textbf{(3-14)}$$

$$d_w,\ d_{qs} \geq 0 \qquad \textbf{(3-15)}$$

where λ_w and λ_{qs} are the electricity and quality prices, and I represents the “budget”.

The solution to the optimization problem is found by forming the Lagrangian, and solving the first order Karush-Kuhn-Tucker conditions, taking into account that d_{qs} depends on d_w according to **(3-9)**. The solution is:

$d_{qs} < \overline{qs}, \quad 0 < d_w < \overline{w}:$

$$d_w = \frac{(1-\beta_2/\beta_1)I + (\beta_2/\beta_1)\lambda_{qs}}{\lambda_w} \qquad \textbf{(3-16)}$$

$$d_{qs} = (\beta_2/\beta_1)(\frac{I}{\lambda_{qs}} - 1)$$

$d_{qs} = \overline{qs}, \quad 0 < d_w < \overline{w}:$

$$d_w = \frac{I - \overline{qs} \cdot \lambda_{qs}}{\lambda_w} \qquad \textbf{(3-17)}$$

$d_{qs} < \overline{qs}, \quad d_w = \overline{w}:$

$$d_{qs} = \frac{I - \overline{w} \cdot \lambda_w}{\lambda_{qs}} \qquad \textbf{(3-18)}$$

$d_{qs} < \overline{qs}, \quad d_w = 0:$

$$d_{qs} = \frac{I}{\lambda_{qs}} \qquad \textbf{(3-19)}$$

$d_{qs} = \overline{qs}, \quad d_w = 0:$

$$\text{only solvable if } \lambda_{qs} = \frac{I}{\overline{qs}} \qquad \textbf{(3-20)}$$

Naturally, **(3-19)** and **(3-20)** are only of academic interest. The results do not conform with the normal properties of the Cobb-Douglas function that the proportion of income used on each good is independent of the price ration (cf. [3.22], page 126). The reason for this is that there is a mutual dependency between d_w and d_{qs}.

3.4.3 Characteristics of the tw o-good utility function

To get an idea of the implications of using the model from 3.4.2, it is useful to look at the behaviour of the optimal values for d_w and d_{qs} as a function of prices for various values of β_1 and β_2.

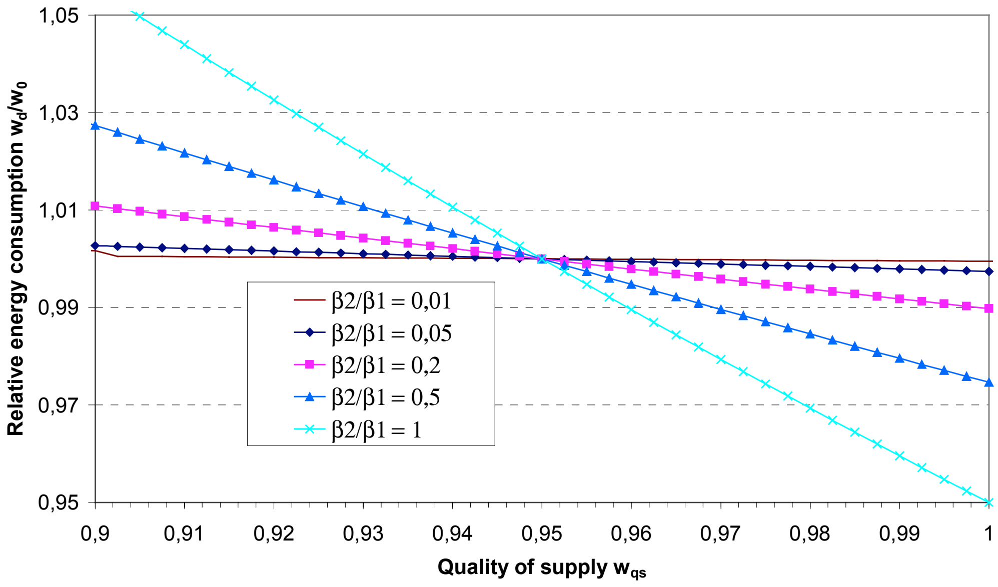

Figure 3-11: Iso-utility curves for different consumers, scaled to an energy consumption of 1.0 for a quality of 0.95.

Figure 3-11 shows iso-utility curves for different consumers, scaled such that their relative electricity consumption is 1.0 for a quality of 0.95[17]. This could be a random rationing situation, where 5 % of demand has to be rationed. A consumer with $\beta_2=\beta_1$ values electricity and quality equally, and will be willing to reduce his electricity consumption with 5 % to obtain a quality of 1.0. However, a consumer with $\beta_2=0.05\cdot\beta_1$ values quality far less, and is hardly willing to reduce consumption to obtain a quality of 1.0

Another way to illustrate the behaviour of the model is to look at optimal energy consumption and quality of supply as a function of the price of quality. Energy is scaled as in **(3-11)**, while the price of quality is taken as a ratio α of the budget I. $\alpha=0$ means that quality is available at no cost, while $\alpha=1$ means that the whole budget must be used on quality to obtain a value of 1.

[17] The curves appear to be straight lines over this short interval, but naturally have the same shape as in
Figure 3-10.

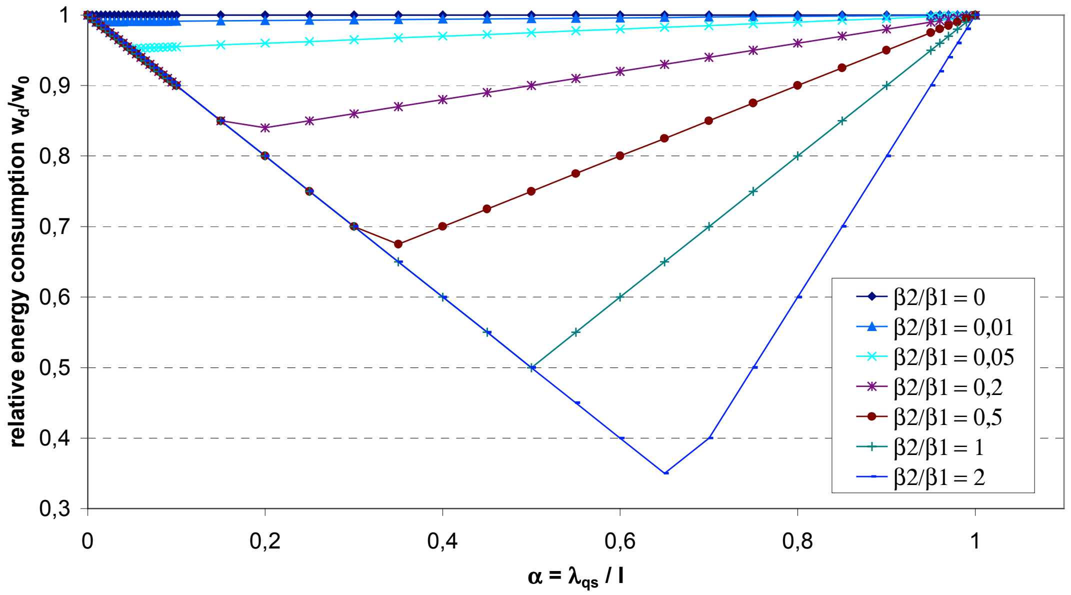

Figure 3-12: Relative optimal energy consumption

The figure shows that optimal electricity consumption falls as a function of α up to a certain point (here **(3-17)** prevails), until maximum quality is no longer optimal, **(3-16)** becomes valid and electricity consumption increases. The point where this happens is given by the value of α where d_{qs} in **(3-16)** is equal to 1:

$$\alpha = \frac{\beta_2/\beta_1}{1+\beta_2/\beta_1} \qquad \textbf{(3-21)}$$

For consumers with small values of β_2/β_1, i.e. with a low preference for quality, optimal relative electricity consumption is close to 1.0 regardless the price of quality. Generally, consumers prefer a high electricity consumption in this model as long as the price for quality is low. When the price increases, some electricity consumption is given up to uphold enough quality, but when the price becomes very high, it becomes "unaffordable", and is given up for electricity, because it makes no sense to buy only quality. Similarly, one can describe optimal quality of supply as a function of α. When quality is available at zero cost, naturally all consumers prefer the maximum amount. This is the typical situation under low and medium load operation, when the short-term cost of keeping reserves is very low. At a certain price of quality, depending on β_2/β_1, consumers give up to obtain maximum quality to be able to obtain enough energy. For low values of β_2/β_1 this occurs for relatively low prices, with a sharp drop. For higher values of β_2/β_1, the transistion is more moderate. This is shown in Figure 3-13.

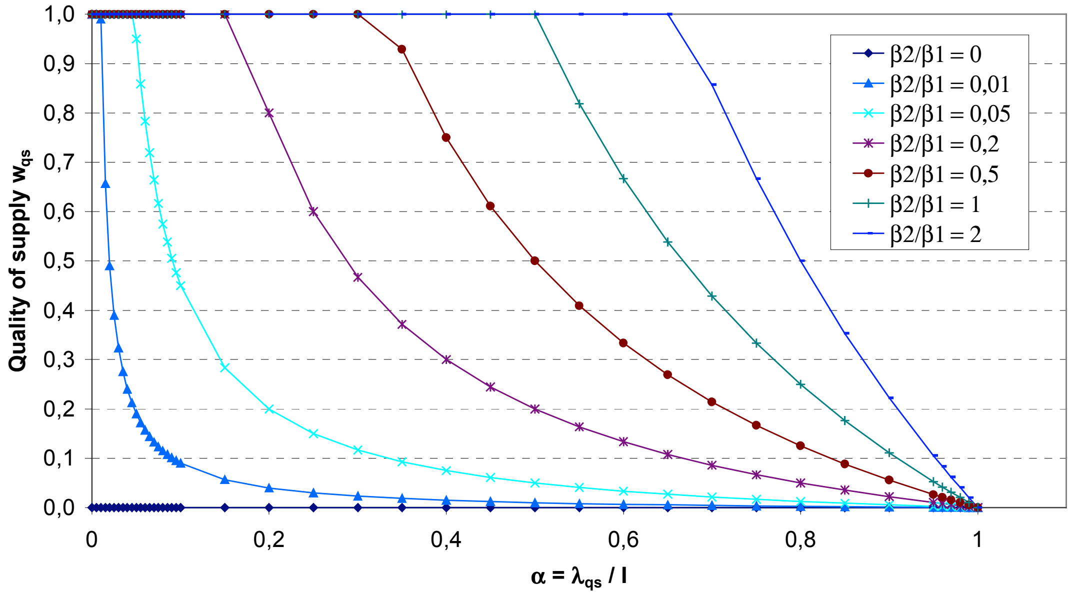

Figure 3-13: Optimal consumption of quality of supply

Under normal operating conditions, the cost of quality is low, and if the price reflects the cost, it seems natural that consumers prefer maximum quality. Under constrained operating conditions, if the price of quality would reflect the long run cost of peaking capacity that is used only a few hours annually, the price of quality may well be represented by values of α between 0.5 and 1.0. In this case there will probably be large variations between consumers' preferences for electricity and quality. Flexibility in the market should be such that these variations are exploited.

3.5 The demand side role in restructured power systems

In the first section of this chapter, a review of electricity pricing theory has been given. It was demonstrated that there has been disagreement about the correct way to price peak demand. On one side stand the proponents of marginal cost pricing – on the other the defenders of an additional mark-up to ensure the coverage of capital costs. The whole discussion is blurred by the "need to keep constant prices", which has technological, economic and socio-psychological ("people do not like price variations") reasons.

The ultimate reason to have peaking capacity available is to be able to supply electricity with high reliability also during peak hours. Traditionally, engineering methods have been employed to design the generation system in such a way that expected peak demand could be covered. In an energy-only market, this is not possible.

In Section 3.3, a theoretical argument has been built around the role of consumers. It has been shown that the kernel of the problem lies in the fact that there *is* a capacity limit, but that consumers do not see it because they are not faced with short-term prices. If no other measures are taken, it can become necessary to resort to random rationing, which is expensive and unacceptable. It was demonstrated that if there is some short-term price elasticity, the problem can be solved by letting the demand side face short-term prices. A simplified example for the Norwegian case illustrates that large savings can be obtained with this approach, even if only a minor part of demand is elastic.

In this context it is appropriate to point out that "profile-based settlement" (presently used in Scandinavia and England and Wales), which allows small consumers to buy electricity from arbitrary suppliers without hourly metering, is detrimental with respect to conveying price signals to consumers. This is especially the case when consumers, as in Norway, are given the option to pay spot prices (albeit with a minor markup). With this form for settlement, the consumer has no incitement to reduce consumption when prices are high, because his own consumption has no impact on the profile. Profile-based settlement is probably a necessary prerequisite to initiate true competition, but should be substituted by metering in the future.

The last part of Section 3.3 prepares the ground for Section 3.4: in the framework so far developed, the effect of letting consumers stand as reserves is analysed. This means they will not have to reduce their consumption *a priori*, but may be switched of on short notice in real time. In Section 3.4 the view is taken that consumers actually have preferences for two goods that are closely related: electricity and quality of supply. A Cobb-Douglas demand function for these goods is developed, and its properties are demonstrated under various assumptions.

It will not be an easy task to ask consumers through surveys what their preferences are for electricity and quality of supply respectively. On the one hand the questions will be hypothetical, and difficult to assess. On the other hand, most consumers in the Western world will view electricity as (nearly) 100 % reliable, and have difficulties with imagining how to handle lower reliability. But the model in Section 3.4 clearly shows that if consumers have such preferences, it will be socially optimal to let them act according to those preferences. In other words, markets must be designed in such a way that this is made possible. In the following chapters these ideas will be pursued further from different viewpoints.

3.6 References

[3.1] Michael A. Crew, Chitrus S. Fernando, Paul R. Kleindorfer, "The Theory of Peak-Load Pricing: A Survey", *Journal of Regulatory Economics*, Vol. 8, No. 3, 1995, pp. 215-248.

[3.2] M. Boiteux, "La Tarification des demandes en pointe: des deplication de la théorie de la vente au coût marginal", *Revue générale de l'électricité*, August 1949 (in French)

[3.3] M. Boiteux, "Peak-load pricing", *Journal of business*, Vol. 33, No. 2, 1960, pp. 157-179.

[3.4] Oliver E. Williamson, "Peak-Load Pricing and Optimal Capacity Under Indivisibility Constraints", *American Economic Review*, Vol. 56, September 1966, pp. 810-827.

[3.5] Ralph Turvey, "Optimal pricing and investment in electricity supply: an essay in welfare economics", London, George Allen & Unwin, 1968.

[3.6] Ralph Turvey, Dennis Anderson, "Electricity Economics. Essays and Case Studies", Published for the World Bank, The John Hopkins Press, 1977.

[3.7] Gardner Brown Jr. M. Bruce Johnson, "Public Utility Pricing and Output Under Risk", *American Economic Review*, Vol. 59, September 1969, pp. 119-128.

[3.8] M.L. Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand: Comment", *American Economic Review*, Vol. 63, March 1973, pp. 224-229.

[3.9] D.W. Carlton, "Peak Load Pricing with Stochastic Demand", *American Economic Review*, Vol 67, December 1977, pp. 1006-1010.

[3.10] D.W. Carlton, "Market Behavior with Demand Uncertainty and Price Inflexibility", *American Economic Review*, Vol 68, September 1978, pp. 571-587.

[3.11] R.A. Meyer, "Monopoly Pricing and Capacity Choice under Uncertainty", *American Economic Review*, Vol 65, June 1975, pp. 426-437.

[3.12] Michael A. Crew and Paul R. Kleindorfer, "Public Utility Economics", MacMillan Press, London, 1979.

[3.13] B.M. Mitchell, W. Manning, J.P. Acton, *Peak-Load Pricing*, Ballinger, Cambridge, MA, 1978.

[3.14] Roland Andersson, Mats Bohman, "Short- and long-run marginal cost pricing. On their alleged equivalence", *Energy Economics*, October 1985.

[3.15] J.A. Kay, "Recent Contributions to the Theory of Marginal Cost Pricing: Some Comments" *The Economic Journal*, Vol. 81, No. 322, June 1971.

[3.16] Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, Roger E. Bohn, "Spot Pricing of Electricity", Kluwer Academic Publishers, 1988.

[3.17] Ross Baldick, R. John Kaye, Felix F. Wu, "Electricity Tariffs Under Imperfect Knowledge of Participant Benefits", *IEEE Transactions on Power Systems*, Vol. 7, No. 4, November 1992, pp. 1471-1482

[3.18] Balbo H. Kim, Martin L. Baughmann, "The Economic Efficiency of Alternatives for Revenue Recollection", *IEEE Transactions on Power Systems*, Vol. 12, No. 3, August 1997, pp. 1129-1135

[3.19] R. John Kaye, Felix F. Wu, Pravin Varaiya, "Pricing for System Security", *IEEE Transactions on Power Systems*, Vol. 10, No. 2, May 1995, pp. 575-583

[3.20] I.J. Pérez-Arriaga, C. Meseguer, "Wholesale Marginal Prices in Competitive Generation Markets", *IEEE Transactions on Power Systems*, Vol. 12, No. 2, May 1988, pp. 710-717

[3.21] X. Vieira Filho et al. "Efficient Pricing Schemes in Competitive Environments Using Cooperative Game Theory", Proceedings of the 13th PSCC, Trondheim, Norway, June 28 – July 2, 1999.

[3.22] Walter Nicholson, "Microeconomic Theory, Basic Principles and Extensions", Dryden Press International Edition, Fifth Edition, 1992

[3.23] David M. Kreps, "A course in Microeconomic Theory", Harvester Wheatsheaf, 1990.

[3.24] "Planleggingsbok for fordelingsnett", EFI, 1993 (in Norwegian).

[3.25] Roy Billinton, Ronald N. Allan, "Reliability Evaluation of Power Systems", Pitman Publishing, 1984

[3.26] "Applied Reliability Assessment in Electric Power Systems", Edited by Roy Billinton, Ronald N. Allan, Luigi Salvaderi, IEEE Press, 1991

[3.27] L.T. Anstine, R.E. Burke, J.E. Casey, R. Holgate, R.S. John, H.G. Stewart, "Application of Probability methods to the Determination of Spinning Reserve Requirements for the Pennsylvania-New Jersey-Maryland Interconnection", *AIEE Transactions on Power Apparatus and Systems*, Vol. 82, October 1963, pp. 726-735, reprinted in [3.26].

[3.28] A.D. Patton, "A probability method for bulk power system security assessment, I – Basic concepts", *AIEE Transactions on Power Apparatus and Systems*, Vol. PAS-91, No. 1, January/February 1972, pp. 54-61, reprinted in [3.26].

[3.29] B.H. Bakken, "Technical and Economic Aspects of operation of Thermal and Hydro Power Systems", *PhD thesis* NTNU, Trondheim, Norway, 1997

[3.30] Eric Hirst, Brendan Kirby, "Unbundling Generation and Transmission services for Competitive Electricity Markets", Oak Ridge National Laboratory, ORNL/CON-454, January 1998

[3.31] John Tschirhart, "On public utility pricing under stochastic demand", *Scottish Journal of Political Economy*, Vol. 27, No. 3, November 1980, pp. 216-234.

[3.32] S. Abraham Ravid, "Reliability and Electricity Pricing", *Journal of Economics and Business*, Vol. 44, 1992, pp. 151-159.

Chapter 4: THE SHORT-TERM, CAPACITY CONSTRAINED EQUILIBRIUM

The short-term operation of power systems is traditionally assessed using Unit Commitment (UC) or dispatch programs with typical time horizons from several hours to one week. Numerous papers have been published on these topics, proposing different solution methods (e.g. Lagrange Relaxation, Mixed Integer Programming, Genetic Algorithms) and including various special problem characteristics like ramping constraints, the inclusion of the transmission network and emission constraints. A literature synopsis of unit commitment can be found in [4.1].

In most of the work, system security is taken implicitly into account through a spinning reserve requirement. The basic assumption is that the availability of reserves reduces the probability of loss of load below an acceptable limit. With inelastic demand and a hard reserve constraint, there is no solution if there is a capacity deficiency.

In Chapter 2, it has been argued that, for various reasons, reserve margins (in the system adequacy sense) will decrease after power system restructuring. In Chapter 3, various demand models were introduced. In this chapter, many of these items will be drawn together in a simulation model that computes the optimal utilization of the generation system combined with different load models during capacity constrained operation. The goal of this chapter is to show that it may be necessary to change the traditional reserve requirements in order to avoid load shedding to satisfy these requirements, and to introduce alternative solutions.

The approach in this chapter is from the system operator point of view. In a market setting, it may be envisaged that bids from market agents substitute costs. In that case, the task of the system operator will be to procure optimally the right amount of reserves from the market, in view of their costs and expected cost savings. Some results are presented in Section 4.4.

4.1 A model for the short-term capacity constrained equilibrium

From the traditional/physical point of view, capacity constrained operation can be defined as operation where it is impossible to serve the entire (inelastic) load while at the same time satisfying the reserve constraints. However, with elastic demand and/or flexible reserve requirements, the concept of "capacity constrained operation" needs a more concise definition: operation is capacity constrained if:

- Demand is inelastic, and available capacity minus reserve requirements is less than demand, or
- Demand is elastic, and the demand curve crosses the supply curve at its vertical part where generation is at its maximum given satisfaction of the reserve constraints, cf. Figure 4-1.

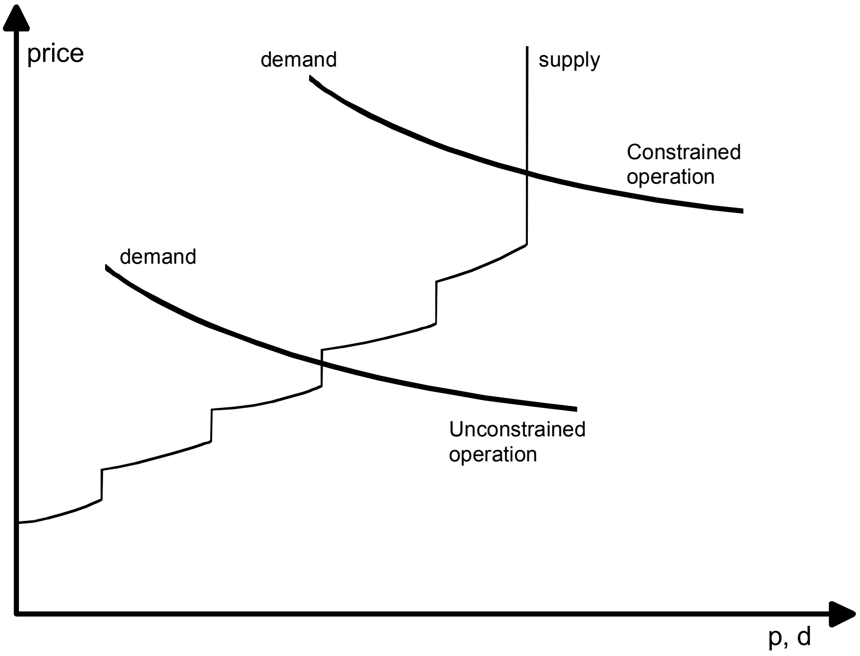

Figure 4-1: Constrained and unconstrained operation

With the demand models introduced in Chapter 3, constrained demand can be further specialized as shown in Figure 4-2.

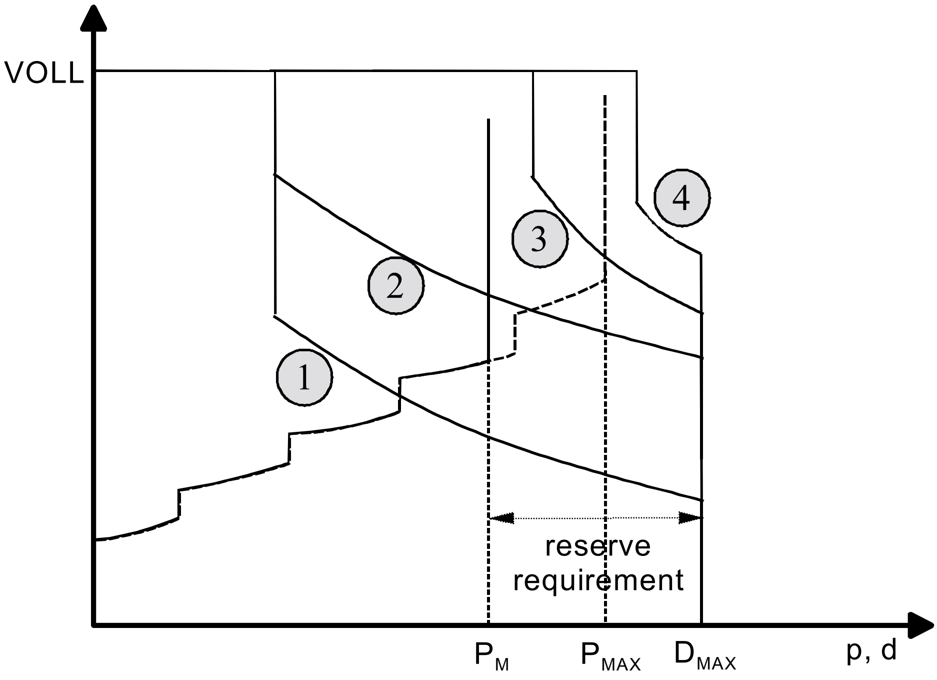

Figure 4-2: Operation with partly elastic demand. 1: unconstrained, 2: constrained, price will reduce demand if elasticity is utilized, 3: constrained, rationing necessary due to reserve requirements, 4: constrained, rationing necessary because demand exceeds available capacity.

The concept of capacity constrained operation can be compared with the definition of peak load in [4.2], Table 1, note a: “The term peak load is used for the case where the multiplier [of the capacity constraint] is strictly positive”.

The aim of the model is to calculate the optimum balance between reserves and generation under capacity constrained conditions, under various assumptions regarding demand flexibility. The approach is comparable with approach taken in [4.3], but focuses more on the situation with a potential capacity deficiency. The time horizon of the analysis is one hour. The objective is to find the minimum cost unit commitment and dispatch, where costs also include the costs of “preventive” load shedding (to satisfy reserve requirements) and the expected cost of loss of load due to generation outages. The base model includes load and generation only, the transmission system is not considered. An extension to a multi-area transport type transmission model is shown in Section 4.2.7.

Conceptually, the analysis consists of two parts: a static UC/dispatch analysis and a dynamic outage analysis. In the first part, the minimum cost dispatch solution is found, while the expected outage cost is calculated in the second part.

Two types of reserves are considered: primary and secondary. It is assumed that primary reserve is frequency controlled and reacts very fast. Thus, as long as the amount of primary reserve is at least as large as outaged generation, no loss of load will occur in the model. The primary reserve contribution of each unit is limited, and a cost is incurred for its provision. Secondary reserve is slower, limited by the units’ ramprates. If not enough primary reserve is available, load has to be shed, and shedding lasts until enough secondary reserve has been activated to restore the balance. To avoid load shedding in all cases, the sum of the primary reserve contributions from all other units must be greater than the largest output from any generator. Naturally, this is supposed always to be the case in large power systems. However, it may be a problem in smaller systems, and it will also be a problem in a large system when there is a capacity deficit.

It is assumed that load is recovered simultaneously with the increase of production by generators that provide reserves. Demand is assumed constant in one hour after an outage, and the loss of load is assumed limited to this hour. Put in another way: all outages are assumed to occur at the beginning of the hour. The concept is illustrated in Figure 4-3. Before the outage, generation equals demand at p_0. When the outage occurs at t=0$^-$, generation drops momentarily, but is assumed to increase quickly due to the effect of primary control at t=0$^+$. From this point, secondary control increases output from available generators until at t=t_r, no more reserves are available, and generation is constant the remainder of the hour (t=T) Lost load is equal to the shaded area in the figure.

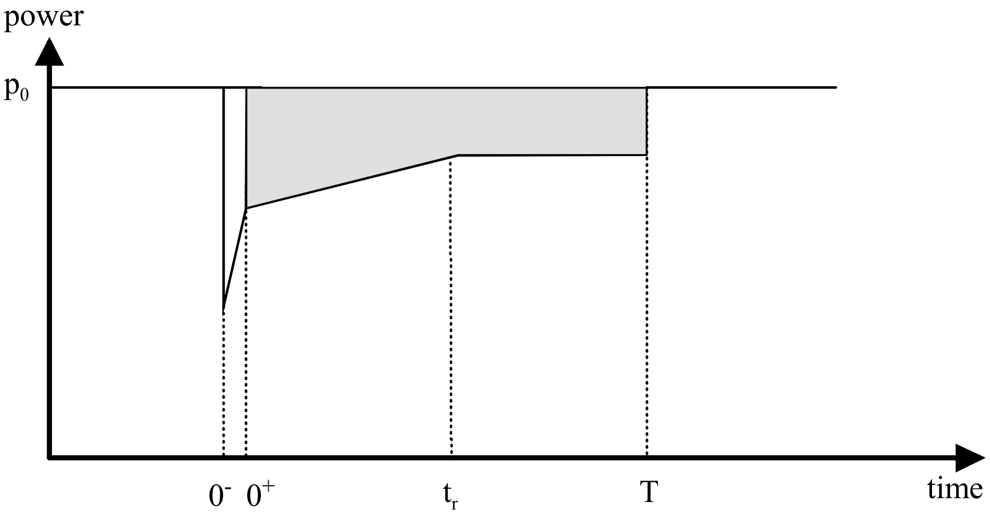

Figure 4-3: Illustration of lost load after generation outage

Generators are modelled with a linear cost function, i.e. a constant marginal cost. A quadratic cost function can be approximated by several linear segments, at the cost of increased computing times. For each generator, a maximum and minimum output is given, i.e. the generator must operate between these limits or not operate at all. This introduces non-convexity to the problem. Furthermore, for each generator a primary reserve cost, ramping rate, start-up time, mean time to failure, start cost and initial status are given.

Several load segments may be defined. For each segment, minimum and maximum demand and corresponding prices are given. Furthermore, the potential contribution to primary and secondary reserve is defined, as well as the activation time for the secondary reserve contribution. The value of lost load and the reduction of this value in the case of early warning are given. Finally, the rationing strategy can be random or optimal, cf. the discussion in Section 3.3.

4.2 Mathematical formulation of the model

Before defining the symbols and the complete model, the modelling of elastic demand will be described in Section 4.2.1

4.2.1 Modelling of elastic demand

Elastic demand is defined by minimum and maximum demand and corresponding prices. A linear function is assumed between these extremes, cf. the left hand part of Figure 4-4.

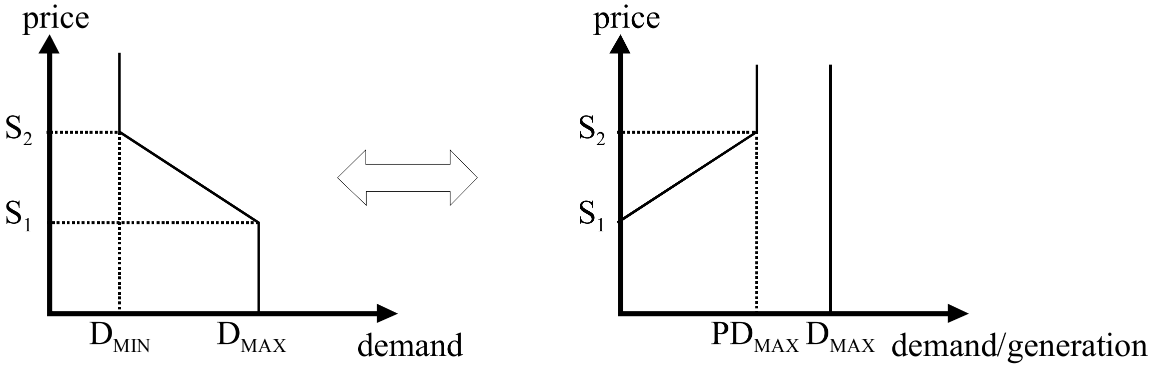

Figure 4-4: Elastic demand modelled as inelastic demand and a generator

To facilitate the handling of elastic demand, it can be modelled as inelastic demand D_{MAX}, combined with a "generator". The output of the "generator" equals the difference between D_{MAX} and actual demand in demand for a given price. In the right hand part of Figure 4-4, the leftmost function represents the "generator", while the vertical line represents inelastic demand. The piecewise linear function in the left hand panel is for any given price equal to the horizontal difference between the two curves in the right hand panel.

The parameters of the generator can be calculated in the following way:

$$PD_{MIN} = 0, \quad PD_{MAX} = D_{MAX} - D_{MIN} \tag{4-1}$$

$$k_D = S_1 + \frac{S_2 - S_1}{D_{MAX} - D_{MIN}} \cdot p \tag{4-2}$$

where p is the production of the "demand generator" and k_D its marginal cost. To see this, compute demand in the left hand panel at a price λ between S_1 and S_2:

$$d = D_{MIN} + \frac{D_{MAX} - D_{MIN}}{S_2 - S_1} \cdot (S_2 - \lambda)$$

In the right hand panel, for the same price λ, generation is such that price equals marginal cost:

$$S_1 + \frac{S_2 - S_1}{D_{MAX} - D_{MIN}} \cdot p = \lambda$$

resulting in

$$p = (\lambda - S_1) \cdot \frac{D_{MAX} - D_{MIN}}{S_2 - S_1}$$

Resulting demand is equal to D_{MAX} - p, which is equal to the expression for d above. Together with the constraints **(4-1)**, inelastic demand D_{MAX} together with the demand generator will thus always result in the same net demand as the elastic demand in the right hand panel of Figure 4-4. The linear marginal cost implies a quadratic cost function, which in the model is represented by a piecewise linear function.

4.2.2 Symbol definitions

Sets:

M_m - The set of outaged generators for outage case *m*

Other sets are not used explicitly in the model description. For example instead of $i \in GEN$ in a summation, only *i* is used, and the set membership is implicit.

Indexes:

i - generator number (i=1..I)

j - demand segment number (j=1..J)

$j1$ - "demand generator" number (jl=1..JL)

k - commodity index, "en": energy, "prim": primary reserve, "sec": secondary reserve

m - outage case number (m=1..M)

t - time period in post-contingency evaluation (t=1..T)

Parameters:

$A0_i$ - Constant term in running-cost function of unit *i* ($)

$A1_{ik}$ - Coefficient of linear term in cost function of unit *i* ($/MW or $/MWh)

AD_{j1} - Linear coefficient in "demand generator" *j*1's cost function ($/MWh)

$DMAX_j$ - Maximum demand for segment *j* (MW)

$DMIN_j$ - Minimum demand for segment *j* (MW)

$DRES_{kj}$ - Maximum contribution to commodity $k\neq$"en" by demand segment *j* (MW)

$INIT_i$ - Initial status (before dispatch) generator *i* (1: spinning, 0: non-spinning)

MPP - Number of minutes per period in post-outage analysis

$PDMAX_{kj1}$ - Maximum production commodity *k*, "demand generator" *j*1 (MW)

$PMAX_{ki}$ - Maximum production commodity *k*, generator *i* (MW)

$PMIN_{ki}$ - Minimum production commodity *k*, generator *i* (MW)

$PMAX0_{ki}$ - Maximum non-spinning production commodity *k*, generator *i* (MW)

PR_m - Probability of outage case *m*

RR_i - Ramping rate generator *i* (MW/min)

$RVOLL$ - Reduced value of lost load (due to early warning) ($/MWh)

$SCOST_i$ - Start-up cost generator *i* ($)[1]

$TACT$ - Activation time demand reserves (min)

$TSTART_i$ - Start-up time non-spinning reserves, generator *i* (min)

$VOLL$ - Value of lost load ($/MWh)

[1] $SCOST_i$ may include the future impact of starting the unit in this hour. This will normally reduce the actual start cost, cf. [4.4].

Variables:

p_{ki} - Pre-contingency production commodity *k*, generator *i* (MW)

u_i - Pre-contingency status generator *i* (1: spinning, 0: non-spinning)

s_i - Start-up status generator *i* (1: starting, 0: not starting)

$shed_j$ - Shedding of load segment *j* to provide reserves (MW)

ppo_{mit} - Post-contingency generation, outage case *m*, generator *i*, time step *t*

pd_{kj1} - Pre-contingency production of commodity *k* by "demand generator" *j*1

$pdpo_{mj1t}$ - Post-contingency generation, outage case *m*, demand generator *j*1, time step *t*

$pdres_{kj}$ - Pre-contingency contribution of reserve from inelastic demand segment *j*

$dred_{mjt}$ - Post-contingency utilization of reserve from inelastic demand segment *j* in outage case m, time step *t*

4.2.3 Mathematical formulation

The mathematical formulation of the model is given below. Equation **(4-3)** gives the objective to minimize over the variables u_i, p_{ki}, s_i, $pd_{en,j1}$, $shed_j$, ppo_{mit}, $pdpo_{mj1t}$, $dred_{mjt}$.

$$\begin{aligned} \text{MIN} \quad & \sum_i \left[A0_i \cdot u_i + \sum_k A1_{ki} \cdot p_{ki} + SCOST_i \cdot s_i \right] + \\ & \sum_{j1} AD_{j1} \cdot pd_{en,j1} + \sum_j RVOLL \cdot shed_j + VOLL \cdot MPP / 60 \cdot \\ & \sum_m PR_m \cdot \sum_t (\sum_j DMAX_j - \sum_i ppo_{mit} - \sum_{j1} pdpo_{mj1t} - \sum_j dred_n \end{aligned} \qquad \textbf{(4-3)}$$

The two first terms describe the operation cost and the direct cost of providing reserves. The third term expresses the start-up cost. The sum $AD_{j1} \cdot pd_{en,j1}$ expresses the lost consumer utility of reducing demand from maximum. Together, the first four terms represent social surplus as the sum of consumer and producer surplus under normal conditions, when the cost of disruptions are neglected. $RVOLL \cdot shed_j$ represents the cost of shedding load segment *j* to be able to have sufficient reserves available. *RVOLL*, the reduced value of lost load is less than *VOLL*, because shedding is warned beforehand when demand is shed to be able to provide sufficient reserves.

In general, the cost of load interruption is given by $VOLL \cdot \int (d(t)-p(t))dt$, where *d(t)* represents uninterrupted demand and *p(t)* generation, cf. Figure 4-3. The last term in parentheses represents this cost. The term with $DMAX_j$ is constant, but when it is included, the objective represents the expected cost directly. The three last terms represent the contribution from generation, elastic demand and inelastic demand respectively (it is assumed that inelastic demand can contribute to reserves, i.e. be switched off in the case of outages). It has to be multiplied with *MPP*/60 because *t* is an index, which multiplied with *MPP* gives a number of minutes. Finally, it is multiplied with the outage probability[2].

Constraints regarding normal operation:

$$\sum_i p_{en,i} + \sum_j shed_j + \sum_{j1} pd_{en,j1} = \sum_j DMAX_j \qquad \textbf{(4-4)}$$

$$p_{ki} - PMAX_{ki} \cdot u_i - PMAX0_{ki} \cdot (1-u_i) \leq 0 \qquad \forall ki \qquad \textbf{(4-5)}$$

$$\sum_k p_{ki} \leq PMAX_{en,i} \qquad \forall i \qquad \textbf{(4-6)}$$

$$p_{en,i} - p_{prim,i} - PMIN_i \cdot u_i \geq 0 \qquad \forall i \qquad \textbf{(4-7)}$$

$$pd_{k,j1} \leq PDMAX_{k,j1} \qquad \forall k, j1 \qquad \textbf{(4-8)}$$

$$\sum_k pd_{k,j1} \leq PDMAX_{en,j1} \qquad \forall j1 \qquad \textbf{(4-9)}$$

$$s_i - u_i + INIT_i \geq 0 \qquad \forall i \qquad \textbf{(4-10)}$$

$$pdres_{kj} \leq DRES_{kj} \qquad \forall k \neq "en", j \qquad \textbf{(4-11)}$$

Equation **(4-4)** gives the balance between generation and demand. **(4-5)** limits maximum output of energy and reserves from both spinning and non-spinning units, while **(4-6)** limits the sum of generation and reserve contributions. **(4-7)** ensures that running units at least produce minimum generation, and at the same time limit primary reserve to the difference between actual and minimum production[3]. **(4-8)**-**(4-9** have a similar function for "demand-generators". **(4-11)** controls the variable s_i, indicating if a unit is started, incurring start-up costs, while **(4-10)** limits the contribution to reserves from inelastic demand.

[2] The loss of load probability *LOLP* is represented by the sum of those PR_m for which $\sum_t (\sum_i \sum_j DMAX_j - ppo_{mit} - \sum_{j1} pdpo_{mj1t} - \sum_j dred_{mjt})$ is positive, i.e. those outages that result in Loss of Load, cf. Section 4.2.9.

[3] In this view, it is assumed that primary reserve is two-sided: it must be possible to use it both for upward and downward regulation. Alternatively, two different services may be used. The Californian market started with the former, but introduced both variants in the course of 1999.

Constraints regarding post-contingency operation:

$$\sum_j shed_j + \sum_i ppo_{mit} + \sum_{j1} pdpo_{mj1t} + \sum_j dred_{mjt} \le \sum_j DMAX_j \qquad \forall mt \qquad \textbf{(4-12)}$$

$$ppo_{mit} - \sum_k p_{ki} \le 0 \qquad \forall mi, t \ne 0 \qquad \textbf{(4-13)}$$

$$pdpo_{mj1t} - \sum_k pd_{kj1} \le 0 \qquad \forall m, j1, t \ne 0 \qquad \textbf{(4-14)}$$

$$dred_{mjt} - \sum_{k \ne "en"} pdres_{kj} \le 0 \qquad \forall mj, t \ne 0 \qquad \textbf{(4-15)}$$

$$ppo_{mi0} - p_{en,i} - p_{prim,i} \le 0 \qquad \forall mi \qquad \textbf{(4-16)}$$

$$pdpo_{mj10} - pd_{en,j1} - pd_{prim,j} \le 0 \qquad \forall m, j1 \qquad \textbf{(4-17)}$$

$$ppo_{mi0} - p_{en,i} \ge 0 \qquad \forall m, i \notin M_m \qquad \textbf{(4-18)}$$

$$pdpo_{mj1,0} - pd_{en,j1} \ge 0 \qquad \forall mj1 \qquad \textbf{(4-19)}$$

$$ppo_{mit} = 0 \qquad \forall m, i \in M_m, t \qquad \textbf{(4-20)}$$

$$pdpo_{mj1t} = pdpo_{mj1,0} \qquad \forall mj1, t < TACT / MPP \qquad \textbf{(4-21)}$$

$$dred_{mjt} = 0 \qquad \forall mjt < TACT / MPP \qquad \textbf{(4-22)}$$

$$ppo_{mit} - PMAX_{en,i} \cdot u_i \le 0 \qquad \forall mi, t < TSTARTi / MPP \qquad \textbf{(4-23)}$$

$$ppo_{mit} - ppo_{mi0} - RR_i \cdot MPP \cdot t \le 0 \quad \forall mi, t < TSTARTi / MPP \qquad \textbf{(4-24)}$$

$$ppo_{mit} - ppo_{mi0} - PMIN_i \cdot (1 - u_i) - RR_i \cdot MPP \cdot (t - TSTART_i \cdot (1 - u_i)) \le 0 \forall mi, t \ge TSTARTi / MPP \qquad \textbf{(4-25)}$$

$$ppo_{mit} - ppo_{mi,t-1} \ge 0 \qquad \forall mi, t > 0 \qquad \textbf{(4-26)}$$

$$pdpo_{mj1t} - pdpo_{mj1,t-1} \ge 0 \qquad \forall mj1, t > 0 \qquad \textbf{(4-27)}$$

$$dred_{mjt} - dred_{mj,t-1} \ge 0 \qquad \forall mj, t > 0 \qquad \textbf{(4-28)}$$

p_{ki}, s_i, $shed_j$, ppo_{mit}, $pd_{k,j1}$, $pdpo_{m,j1,t}$, $dred_{mit}$ $\ge$ 0, u_i $\in\{0,1\}$ $\forall i,j,k,m,j1,t$

Equation **(4-12)** is the equivalent of the balance equation, but in this case generation can be lower than maximum demand, which equals desired consumption under undisturbed operation. **(4-13)** limits post-outage generation ppo_{mit} to the sum of pre-outage generation and reserve availability from each generator. **(4-14)** and **(4-15)** do the same for elastic and inelastic demand, respectively. **(4-16)** and **(4-17)** limit the initial contribution to primary reserve at t=0^+ from each generator respectively demand generator. **(4-18)** and **(4-19)** are not strictly necessary from a cost- or energy-not-served calculation point of view, but they avoid unrealistic solutions if surplus primary reserve is available for a particular outage. **(4-20)** sets the output of outaged generators to zero. **(4-21)** and **(4-22)** limit demand side contribution to zero before its activation time. **(4-23)** limits the output of non-spinning units to zero before they can be started up, while it is redundant for spinning units. **(4-24)** limits the output of spinning units according to their ramprates before their start-up time[4], while **(4-25)** does the same for all units after their startup time. Finally, **(4-26)**-**(4-28)** avoid output reduction after all load has been recovered. Again, this it not necessary from a optimization point of view, but avoids unrealistic solutions.

4.2.4 Price calculation

The primary objective of the model presented here is to investigate how alternative ways to control system security influence obtained security and costs. However, it is also of interest to calculate marginal cost, which to some extent can be seen as an indication of the system price. Only to some extent, because the validity of the dual value of the balance equation **(4-4)** is limited. In a linear model, the calculated dual values are only valid within certain ranges of variations of the right hand side of the constraints. In a mixed linear integer model, the dual values are in addition only valid for the actual integer solution of the problem. Moreover, when the dual value of **(4-4)** is used as price, there is no guarantee that all units recover their startup and noload cost[5]. If an algorithm like this should be used as a practical pricing model, it would be necessary to take certain measures to handle these difficulties. With these limitations, the dual value of **(4-4)** gives an indication of the energy price, and makes it possible to compare the influence of various assumptions on this price.

Because total demand is included in the objective function **(4-3)**, $T \cdot VOLL \cdot MPP/60$ has to be added to the dual value of **(4-4)**. Moreover, the right hand side of **(4-12)** has to be substituted with the left hand side of **(4-4)** to ensure that demand appears only once in the right hand side.

The price reflects the increase in total cost caused by a marginal increase in demand. This price increase has two components: the increase in generation cost and the increase in the cost of energy not served or a security cost. The latter is equivalent to the England and Wales capacity charge, but instead of being added after calculation of the marginal price, it is an integrated part of the price calculation.

[4] Start-up time is irrelevant for spinning generators, but if **(4-24)** had been valid for all time steps, the equations after $TSTART_i$ would have been redundant due to **(4-25)**.

[5] A model that does take these costs into account is introduced in Chapter 5.

4.2.5 Reserve requirements

The formulation of the model in Section 4.2.3 implies cost minimization, i.e. resulting system security depends primarily on generation costs, outage probabilities and consequences and the value of lost load. Normally, system security is handled by imposing certain requirements. In the model, this can be accommodated by adding appropriate constraints:

$$\sum_i p_{prim,i} \geq R_{prim}, \quad \sum_i p_{sec,i} \geq R_{sec} \qquad \textbf{(4-29)}$$

or

$$\sum_{k \neq "en"} \sum_i p_{ki} \geq R \qquad \textbf{(4-30)}$$

The former constrains both types of reserves to minimum values, a normal practice in power systems. However, it is inefficient, and should rather be formulated as:

$$\sum_i p_{prim,i} \geq R_{prim}, \quad \sum_i (\, p_{prim,i} + p_{sec,i}\,) \geq R_{prim} + R_{sec}$$

(4-30) only constrains the sum of reserves, leaving it to the optimization to divide between primary and secondary reserves.

4.2.6 Load uncertainty

Load uncertainty can be modelled with the techniques described in [4.5]. The load is described by its expected value and standard deviation, and a normal distribution is assumed. The normal distribution is then approximated by a number of discrete steps.

It is assumed that unit commitment is decided before the load uncertainty is revealed, i.e. the variables u_i and s_i are independent of the uncertainty state h. The other variables receive the extra index h to indicate the uncertainty state, e.g. p_{hki}. Naturally, the model becomes very large if the number of states is large. If load uncertainty is small, it should not be necessary to use many states. If it is large, the number of states should be balanced against the number of outage cases, i.e. a number of cases with obviously small impacts could be left out of the calculations.

4.2.7 Multi-area modelling

The model described so far is a single-area model. It is assumed that tie line constraints have no vital impact on the solution, because focus is on generation capacity. However, in real systems, transmission constraints often have impact on unit commitment, dispatch and reserve allocation.

A simple transport-model version of inter-area tie lines can be included in the model. In this case, the following additional model attributes must be defined:

Sets:

AREA - The set of areas
LINES - The set of lines
PA_a - The set of generators in area *a*
DA_a - The set of demand segments in area *a*

Indexes:

a - Area index

Parameters:

$LINECAP_{ab}$ - Line capacity for line from area *a* to area *b* (MW)

Variables:

pl_{ab} - Flow on line from area *a* to area *b* (MW)
$plpo_{mabt}$ - Post-contingency flow from *a* to *b* for outage case *m*, period *t* (MW)

Constraints:

The balance equation **(4-4)** becomes:

$$\begin{aligned} &\sum_{i\in PA_a} p_{en,i} + \sum_{j\in DA_a} shed_j + \sum_{j1\in DA_a} pd_{en,j1} - \sum_{ab} pl_{ab} + \sum_{ba} pl_{ba} = \\ &\sum_{j\in DA_a} DMAX_j \qquad \forall a \end{aligned} \qquad \textbf{(4-31)}$$

A similar change is made to the inequality **(4-12)**. Moreover, the following bounds are added:

$$\begin{aligned} &pl_{ab} \le LINECAP_{ab} && \forall ab \\ &pl_{ab} \ge -LINECAP_{ab} && \forall ab \end{aligned} \qquad \textbf{(4-32)}$$

and analogous for $plpo_{ba}$.

4.2.8 The possibility of system collapse

A central assumption in the model is that lost load can be retrieved during the activation of secondary reserves. This assumption limits the damage of outages, and reduces reserve requirements. However, during heavy load conditions, outage of a large generator may not just necessitate limited shedding of demand, but may jeopardize system security to the extent that the probability of system collapse becomes unacceptably high. If this possibility is taken into account, it seems plausible that higher reserve levels are required.

A similar approach is taken in [4.6], where the cost of system collapse is included in the objective function that is to be minimized. Kaye et al. differentiate between *survivable* *contingencies* and *system collapse*. The former can be recovered by using participants' "contingency offerings", while the latter entail a K_{out} multiplied with the probability of collapse.

To explore this effect in the present context, a potential probability of system collapse $MCOLPRO_i$ was attached to each single outage. In the test calculations, tentative values were used. Realistic estimates could be based on combinations of simulations and expert judgement.

The actual probability of system collapse, given the outage of unit *i* is given as:

$$colpro_m \geq \frac{MCOLPRO_m}{100} \cdot \frac{p_{en,m} - \sum\limits_{i \neq m} p_{prim,i}}{PMAX_{en,m}} \qquad \textbf{(4-33)}$$

Thus, if available primary reserves exceed the capacity of unit *m*, the rightmost expression is less than zero, and the outage does not contribute to the probability of system collapse (because $colpro_m$ will be zero in the optimum because of the general non-negativity requirement). If, on the other hand the capacity exceeds available primary reserves, the contribution of the outage to the probability of system collapse depends on the ratio between primary reserves and the unit's capacity. Figure 4-5 shows the probability of system collapse as a function of the pre-contingency output p_m of generator *m* for two different generator sizes.

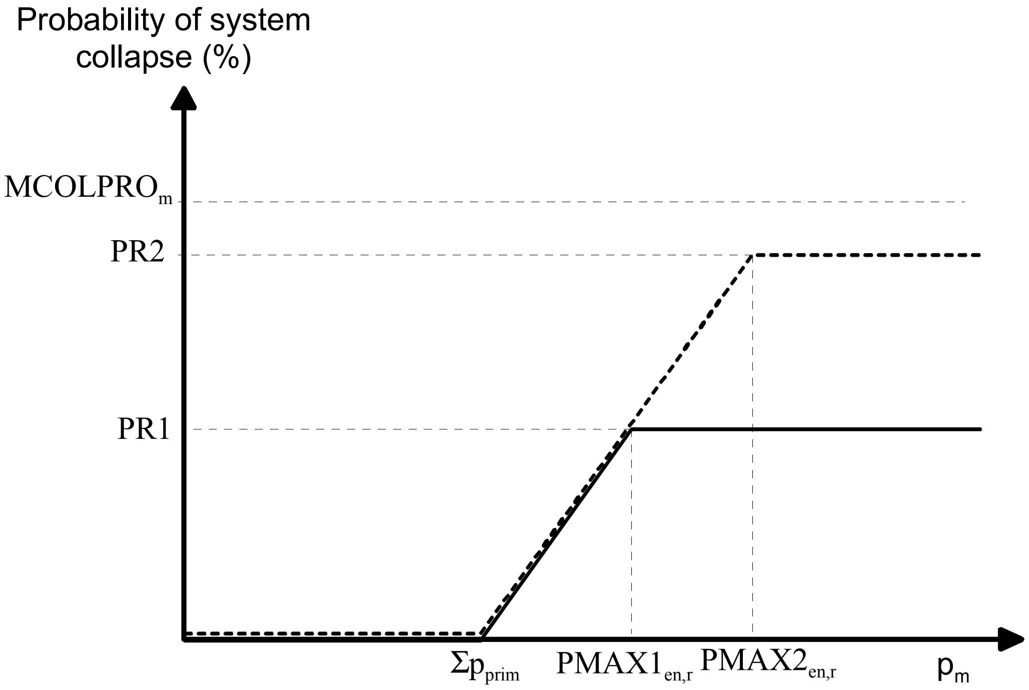

Figure 4-5: Probability of system collapse as a function of pre-contingency generation p_m

For multiple outages, **(4-33)** is replaced by:

$$colpro_m \geq \frac{min(\sum_{r \in M_m} MCOLPRO_r ,100)}{100} \cdot \frac{\sum_{r \in M_m} p_{en,r} - \sum_{i \notin M_m} p_{prim,i}}{\sum_{r \in M_m} PMAX_{en,r}} \qquad \textbf{(4-34)}$$

Furthermore, it is assumed that a system collapse has the effect that all demand is shed for exactly one hour, i.e. the following term is added to the objective function **(4-3)**: $\sum_m VOLL \cdot PR_m \cdot colpro_m \cdot \sum_j DMAX_j$.

Naturally, this is a very simplified version of reality. Its main function is to explore the results of increased consideration to system security or increased risk aversion.

4.2.9 Constraints on *EENS*, *LOLP* and the probability of system collapse

As an alternative to either cost minimization taking into account outage costs or reserve requirements, it is possible to constrain the value of either Expected Energy not Served, *EENS*, the Loss of Load Probability, *LOLP* or the Probability of System Collapse (*POSC*) directly.

For (relative) *EENS* this can is obtained by adding the inequality **(4-35)**:

$$\begin{aligned} & MPP/60 \cdot \sum_m PR_m \cdot (\sum_j (DMAX_j - shed_j) - \sum_{j1} pd_{en,j1} - \\ & \sum_t (\sum_i ppo_{mit} - \sum_{j1} pdpo_{mj1t} - \sum_j dred_{mjt}) \leq \\ & MAXEENS \cdot \sum_j \left(DMAX_j - shed_j \right) - \sum_{j1} pd_{j1} \end{aligned} \qquad \textbf{(4-35)}$$

The left-hand side of the inequality represents *EENS*, while the right hand side is equal to the maximum allowed value of *EENS* multiplied with pre-contingency demand.

For *POSC* the constraint can be represented by:

$$\sum_m PR_m \cdot colpro_m \leq MAXSYSCOL \qquad \textbf{(4-36)}$$

For *LOLP*, the procedure is slightly more complicated. Outage of a unit contributes to *LOLP* if the unit's generation before the outage exceeds available primary reserves. To model this, a binary variable *lolpcase_m* is defined, which must satisfy the constraint:

$$lolpcase_m \geq \frac{p_{en,m} - \sum_{i \neq m} p_{prim.i}}{PMAX_{en,m}} \qquad \forall m \qquad \textbf{(4-37)}$$

If primary reserves exceed the generation of unit *m*, the numerator of the right hand side is negative, and *lolpcase*$_m$ becomes zero. In the opposite case, the right hand side becomes positive, but will never exceed 1, thus the binary variable *lolpcase*$_m$ becomes one. The following inequality limits *LOLP* to its maximum allowed value:

$$\sum_m PR_m \cdot lolpcase_m \le MAXLOLP \qquad \textbf{(4-38)}$$

4.2.10 Model solution

The described model is linear, but contains the binary variables u_i and s_i (and if *LOLP* is constrained, *lolpcase*$_m$). This suggests the use of a mixed integer linear program (e.g. [4.7]). The use of a branch and bound procedure can be expected to be reasonably effective, because focus is on situations where capacity is constrained, which means that many of the available units will have to be committed to obtain a feasible solution. The number of nodes in the binary tree will therefore not become extremely high.

The model was solved with the modelling language AMPL [4.8] using CPLEX [4.9] as a solver[6].

4.3 Simulations with fixed r eserve requirements

In the calculations in this section and Section 4.4, available capacity exceeds demand, but is in a number of cases insufficient to cover traditional reserve requirements. Each simulation is run with a time horizon of one hour. All outages with a probability greater than 10^{-4} are assessed as part of the optimization, cf. the mathematical description in the previous paragraph. Each outage is assumed to occur at the beginning of the hour. The objective of the simulations is to minimize total costs, including expected outages costs.

Reserve requirements corresponding to a loss of load probability less than 10^{-3} were calculated using the PJM method, cf. Appendix II. Without load uncertainty, the resulting reserve for the IEEE Reliability Test System (RTS-96) was 402 MW[7]. In this chapter, results are presented for a fixed reserve of 400 MW. This is implemented in the model by adding the constraint:

[6] It is possible to structure the problem as a Benders decomposition, with the commitment/dispatch problem as the master problem and the contingency analysis as subproblems. Instead of one large problem, this will create two small problems, which can be solved quickly. This approach was tested, but the number of iterations was so high that computing time by far exceeded the single-problem approach. Partly this was due to the interaction between AMPL and CPLEX, which introduces some extra overhead in the iteration process. The approach might be viable for very large problems, with use of CPLEX from e.g. a C program, but it is doubtful if it will be more effective.

[7] With a standard deviation in the load forecast of 2 %, the required reserve margin for this system is 438 MW, and with 4 % 525 MW. For more details, cf. Appendix A.4.2, [4.10] and [4.11]

$$\sum_{i} \sum_{k \notin "en"} p_{ki} = R \tag{4-39}$$

The division between primary and secondary reserves is a result of the model optimization.

4.3.1 Inelastic demand

Simulations were run for the case with inelastic demand to evaluate the model's behaviour and as a reference for the subsequent analyses. The main purpose of these calculations is to explore the characteristics of the model under varying assumptions. The following cases were considered:

Table 4-1: Simulation cases

case	description
base case	cf. Appendix II
noprimcost	the direct cost of primary reserves ($A1_{prim}$) is zero
nostartcost	start-up costs are zero, e.g. because all units are running initially
nosecondary	secondary reserves cannot be used to reduce loss of load
syscollapse	probability of system collapse is attatched to each outage case
prim –50%	primary reserve contribution from each unit is reduced from 5 to 2.5 %
prim +100%	primary reserve contribution from each unit is increased from 5 to 10 %
sec –50%	secondary reserve contribution from each unit is reduced from 20 to 10 %, none from units 1-9
coldres	units 1-9 can contribute to secondary reserve without being committed
resp 10 min	response time of secondary reserve is reduced from 20 to 10 minutes
voll 6000	value of loss of load is increased from 3000 $/MWh to 6000 $/MWh

It might be argued that the way secondary reserve is used to relieve load shedding illustrated in Figure 4-3 is unrealistic. The motivation for the cases *noprimcost*, *nostartcost*, *nosecondary* and *syscollapse* is to evaluate the optimal amount of primary reserve under various assumptions.

Only single outages were considered. It appeared that the inclusion of outages with a probability greater than 10^{-6} (instead of 10^{-4}, an additional 119 cases) did not change optimal commitment and dispatch. Increased expected energy not served was less than 2 %, but had a considerable effect on computation time, which was not considered worth while. Results are shown in the table below:

Table 4-2: Simulation results, fixed reserve 400 MW, inelastic demand 2850 MW

case	primary / secondary reserve (MW)	preven-tive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interrup-tion cost ($)	total cost ($)
base	59/341	0.0	0.1807 ($6.3 \cdot 10^{-5}$)	33621	542	34163
noprimcost	108/291	0.0	0.1199 ($4.2 \cdot 10^{-5}$)	33258	360	33618
nostartcost	59/341	0.0	0.1604 ($5.6 \cdot 10^{-5}$)	30104	481	30585
nosecondary	162/ -	0.0	0.5771 ($2.0 \cdot 10^{-4}$)	34740	1731	36471
syscollaps	163/ 237	0.0	0.0788 ($2.8 \cdot 10^{-4}$)	34740	2615	37355
prim –50%	44/356	0.0	0.2089 ($7.3 \cdot 10^{-5}$)	33776	627	34403
prim +100%	177/223	0.0	0.0721 ($2.5 \cdot 10^{-5}$)	32979	216	33195
sec –50%	123/277	0.0	0.0893 ($3.1 \cdot 10^{-5}$)	35606	268	35874
coldres	50/350	0.0	0.2007 ($7.0 \cdot 10^{-5}$)	30561	602	31163
resp 10 min	123/276	0.0	0.0882 ($3.1 \cdot 10^{-5}$)	34884	265	35149
voll 6000	100/300	0.0	0.1272 ($4.5 \cdot 10^{-5}$)	33862	763	34625

It appears that there is a positive probability of loss of load for all cases. This is because there is not enough primary reserve available in the system to compensate for the loss of the largest units. When the direct cost of primary reserve is zero (*noprimcost)*, optimal primary reserve is almost doubled, resulting in 30 % reduction in *EENS*. Still, the level of primary reserve is not limited by the unit's ability to provide it because of the related opportunity cost: for example, the hydro units do not provide primary reserve because this will increase total fuel costs. Zero start-up costs (*nostartcost*) do not change the optimal level of primary reserve. However, in this case units 1-5 are started up, resulting in faster secondary reserves and a reduction in *EENS*. The cases *nosecondary* and *syscollaps* both result in a much higher level of primary reserve, though still slightly under the maximum amount of 170 MW. Naturally both generation and *EENS* costs are higher in these cases.

A 50 % reduction in the units' ability to provide primary reserves naturally increases *EENS* and total costs, with the opposite result for a 100 % increase. A reduction in the ability to provide secondary reserves, reduces *EENS* due to the increase in primary reserves (because of the requirement of 400 MW total reserves). An interesting observation is that the use of the smaller units 1-9 as non-spinning secondary reserves results in large decrease in generation cost. In this case, none of these units is dispatched, reducing start-up- and zero-load costs. However, *EENS* increases. Decreasing the response time of secondary reserve or increasing VOLL reduces *EENS* and increases generation costs.

Figure 4-6 shows the calculated prices for the various cases. For most cases the price is 21.8 $/MWh, of which 21.6 $/MWh represents the marginal generation cost (which is equal to the variable cost of the marginal plant) and 0.2 $/MWh the security cost. When secondary reserves cannot be used to reduce loss of load, or when the probability of system collapse is taken into account, the generation component rises to 27.3 $MWh, because an increase in generation involves a redispatch from cheaper to more expensive units. However, due to redispatch, security increases with increasing demand, resulting in a negative security cost.

When the primary reserve contribution from each unit is increased to 10 %, a marginal demand increase does not influence security, and the security cost is zero. Finally, an increase in *VOLL* results in a corresponding increase in the security cost.

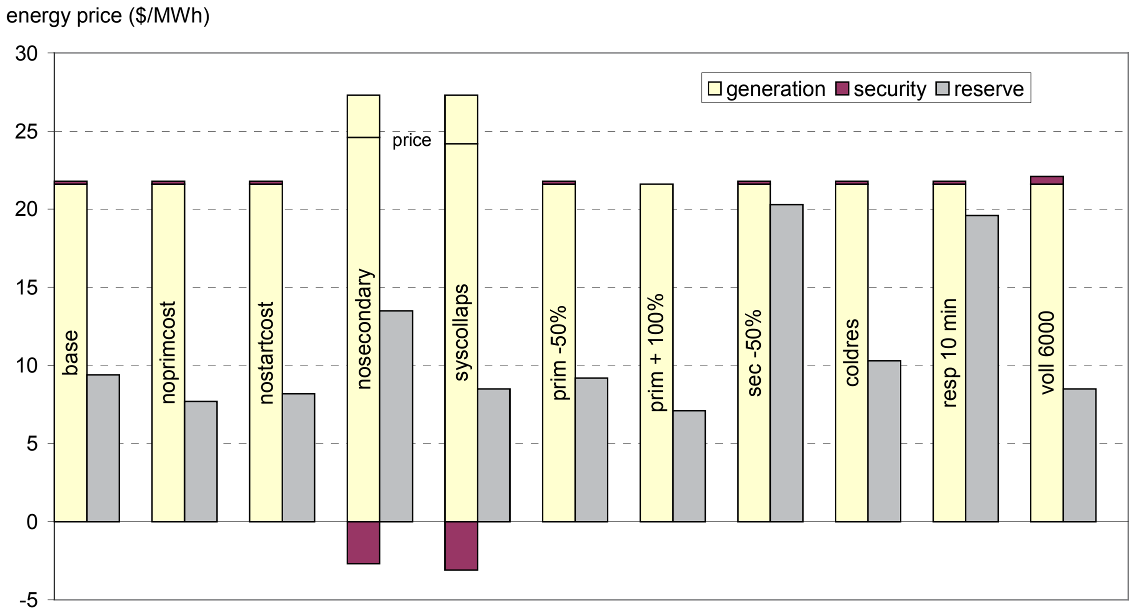

Figure 4-6: Simulated prices, inelastic demand 2850 MW

Figure 4-6 also shows the simulated reserve prices. In most cases they lie between 7 and 10 $/MW, representing the cost of increasing generation on a more expensive unit and decreasing on a less expensive unit to obtain more reserves. In the cases where there is less secondary reserve available, it is necessary to use zero marginal cost hydro units to provide secondary reserves. An increase in the reserve requirement then results in reducing output on a hydro unit, while increasing an expensive thermal unit.

The following table shows the corresponding results for a 10 % load increase to 3135 MW.

Table 4-3: Simulation results, fixed reserve 400 MW, inelastic demand 3135 MW

case	primary / secondary reserve (MW)	preventive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interruption cost ($)	total cost ($)
base	59/341	130.0	0.1637 ($5.2\cdot10^{-5}$)	41870	273491	315361
prim –50%	29/371	130.0	0.2103 ($6.7\cdot10^{-5}$)	41942	273631	315573
prim +100%	177/223	130.0	0.0706 ($2.3\cdot10^{-5}$)	41366	273212	314578
sec –50%	130/270	130.0	0.0847 ($2.7\cdot10^{-5}$)	44999	273254	318253
coldres	49/351	130.0	0.2094 ($7.0\cdot10^{-5}$)	35472	273629	309101
resp 10 min	86/314	130.0	0.1226 ($3.9\cdot10^{-5}$)	42506	273368	315874
voll 6000	80/320	130.0	0.1393 ($4.4\cdot10^{-5}$)	41989	546835	588824

In this case, there is a capacity deficiency. With inelastic load and fixed reserve requirements, the only way to solve this is by preventive load shedding to satisfy the reserve requirements. The results are similar to those in the previous table. Actually, the relative *EENS* is slightly lower in this case, due to the fact that in this case all units are dispatched, which gives a faster overall rate of output increase after an outage. The exception is the case with non-spinning reserves, where units 1-9 provide reserves only, resulting in a low generation cost but a high *EENS*. The load interruption cost includes the cost of preventive shedding.

In this case, the price equals the reduced value of lost load *RVOLL* = 2100 $/MWh, because an increase in demand will automatically increase preventive shedding correspondingly. Reserve prices are close to 2100 $/MW, because an increase in the reserve requirement will increase preventive shedding. At the same time, the expected cost of energy not served is slightly reduced in case of outages.

4.3.2 Elastic demand

It is now assumed that demand is inelastic for "normal" prices, but elastic when prices increase due to a capacity shortage. The demand model of Figure 4-4 is used with S_1 = 50 $/MWh S_2 = *RVOLL* = 2100 $/MWh and demand D_{MAX} equal to 3135 MW. Simulations were done for several levels of the elasticity limit D_{MIN}, (P_1 in the context of Section 3.3). If the optimal commitment/dispatch solutions independent of $\mathrm{D_{MIN}}$, the calculation method of Section 3.3.4 can be used directly, but it is not possible to know this beforehand.

In this case, the commitment/dispatch solution appears to be independent of D_{MIN}, i.e. equal to the base case solution in Table 4-3, with the exception of the load interruption cost and total cost. Based on the theory in 3.3, the following costs were calculated:

Table 4-4: Cost of preventive shedding with perfect rationing as a function of elasticity limit D_{MIN} in IEEE-RTS

elasticity limit	0	627	1254	1881	2508	3005	3135
shedding cost	14451	16439	19752	26378	46257	198250	273000

These calculations were verified with simulations that showed almost the same results, with small deviations due to the discretization of elastic demand.

4.3.3 Seemingly inelastic demand

If demand is elastic, like in the previous section, but this elasticity is invisible for the reasons discussed in Section 3.3, random instead of perfect rationing has to be used. The resulting costs are shown in Table 4-5.

Table 4-5: Cost of preventive shedding with random rationing as a function of elasticity limit in IEEE-RTS

elasticity limit	0	627	1254	1881	2508	3005	3135
shedding cost	198250	213200	228150	243100	258050	269900	273000

The results of these calculations are shown in the next figure, which is similar to the corresponding figure in Section 3.3.

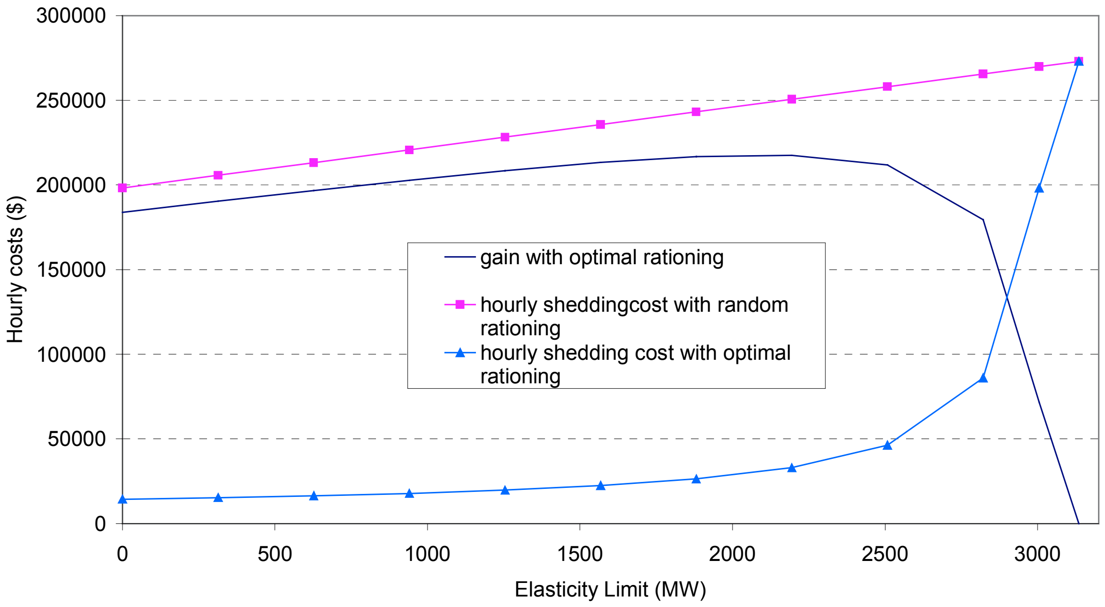

Figure 4-7: Cost of optimal and random preventive shedding as a function of the elasticity limit

4.3.4 Demand side reserves

As discussed in Section 3.4, an alternative to preventive shedding to obtain a specific reserve level is to use demand as reserves. It is assumed that the variable cost of this is zero, but that contributing consumers are paid for being available as reserve. It is assumed that demand only can contribute to secondary reserves, and with the exception of the last simulation, an activation time of 10 minutes is used. Table 4-6 summarizes the results for these simulations. As a reference, the base case results from Table 4-3 are also shown.

Table 4-6: Simulation results, fixed reserve 400 MW, demand side reserves

demand side secondary reserve (MW)	generation primary / secondary reserve (MW)	preven-tive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interrup-tion cost ($)	total cost ($)
0 (base case)	59/341	130.0	0.1637 ($5.2 \cdot 10^{-5}$)	41870	273491	315361
50	50/300	80.0	0.1851 ($5.9 \cdot 10^{-5}$)	42396	168556	210952
100	44/256	30.0	0.2056 ($6.6 \cdot 10^{-5}$)	42989	63616	106605
150	41/209	0.0	0.2210 ($7.0 \cdot 10^{-5}$)	42608	662	43270
100 / 30 min	44/256	30.0	0.2857 ($9.1 \cdot 10^{-5}$)	42989	63857	106846

The opportunity to use demand side reserves leads to considerable cost reductions, net of the cost of installing necessary technology to make such a solution feasible. Because demand only contributes with secondary reserves, which take some time to activate, *EENS* increases compared with the base case solution. Generation costs are also higher, because production is higher than in the base case. In the cases where preventive load shedding is necessary, the price is 2100 $/MWh, like in Section 90. The only exception is the case with 150 MW demand side reserves, in which case the price is 22.0 $ MWh.

4.4 Simulations with optimal reserve requirements

In the previous section, fixed reserve requirements were imposed, which were produced in an optimal manner with respect to a number of outage situations. In this section, reserve amounts are a result of the optimization process, under various assumptions. Especially interesting are the results in Sections 4.4.3 and 4.4.4, where constraints are imposed directly on *EENS* and *LOLP* respectively.

4.4.1 Cost minimization without additional requirements

In this section, optimal reserve amounts are the result of the cost optimization. Outages are taken into account through their impact on expected cost of loss of load. The results are shown in the next table:

Table 4-7: Simulation results, optimal reserve

case	primary / secondary reserve (MW)	preven-tive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interrup-tion cost ($)	total cost ($)
2850 MW	34/164	0.0	0.5220 ($1.8 \cdot 10^{-4}$)	28323	1566	29889
3135 MW	44/85	0.0	0.7583 ($2.4 \cdot 10^{-4}$)	37730	2275	40005

Compared with Table 4-2 and Table 4-3, the reserve level has been considerably reduced, resulting in an increase in *EENS*. However, this is more than compensated by reduced generation costs, and in the 3135 MW case, the reduction in the cost of preventive shedding. The main reason for the difference in generation cost between these simulations and the corresponding simulations with fixed reserves is that none of the 12 and 20 MW oil units is started up. Their start-up respectively standby cost is too high to compensate for their contribution to the reduction of *EENS*. This contribution is very limited, because of their low ability to provide primary reserve (5 % or 0.6 and 1.0 MW respectively). Prices (energy + security component) are 21.6 + 1.0 = 22.6 and 21.6 + 15.5 = 37.1 \$/MWh for the 2850 and 3135 MW demand cases respectively, illustrating the increase in the cost of security when demand rises.

Because preventive load shedding is unprofitable under these assumptions, elasticity of demand as defined in Section 4.3.2 does not change these results, as long as the lower price limit S_l is above the highest marginal generation cost

4.4.2 The cost of system collapse

As pointed out in Section 4.3.1, the results may be too optimistic if the probability of system collapse is not taken into account. Results are shown in Table 4-8 for the cost function described in Section 4.2.8.

Table 4-8: Simulation results, consideration to the probability of system collapse

case	primary / secondary reserve (MW)	preventive shedding (MW)	expected energy not served (GWh / relative)	generation cost (\$)	load interruption cost (\$)	total cost (\$)
2850 MW	153/61	0.0	0.3377 ($1.2{\cdot}10^{-4}$)	30328	3832	34161
3135 MW	130/0	0.0	0.6653 ($2.1{\cdot}10^{-4}$)	39251	6200	45451
3135-2 hour	165/5	0.0	0.4422 ($1.4{\cdot}10^{-4}$)	41632	7402	49034
3135-4 hour	165/5	0.0	0.4422 ($1.4{\cdot}10^{-4}$)	41632	13477	55109
3135-coldres	130/140	0.0	0.3145 ($1.0{\cdot}10^{-4}$)	39251	5147	44398

The first two rows give the results for the 2850 and 3135 MW cases, where a blackout time of 1 hour is assumed. The next two rows show effect of increasing the blackout time to 2 and 4 hours respectively. In the first two cases, still none of the 12 and 20 MW units is started up. In the last two cases, two of the 20 MW units are used.

The last row shows the result of allowing secondary reserve contribution from the non-spinning units 1-9. In this case, it is optimal to utilize these reserves completely, which is logical because it involves no cost. It gives a considerable reduction in *EENS* from outages not resulting in system collapse.

Figure 4-8 shows the simulated prices for these cases. For the 2850 MW case, the result is straightforward – the possibility of system collapse results in a considerable price increase.

For the 3135 MW case with a 1-hour blackout duration and the case where cold reserves can contribute, the generation component is actually negative. This is because output is increased on the nuclear units, which have a marginal cost of zero, while primary reserves are reduced, reducing total generation cost. The lower security cost for the case where cold reserves can contribute is caused by the fact that the optimal secondary reserve is much higher in this case. The results for the 2- and 4-hour blackout duration are more intuitive, but it seems strange that the price is lower for the 2-hour case than for the 1-hour case. The reason is that the 2- and 4-hour cases have other commitment solutions than the 1-hour case. This illustrates the difficulties with using this kind of algorithms for pricing: there is no monotonous relationship between demand and price. Generally, it is clear that including the possibility of system collapse increases the price considerably.

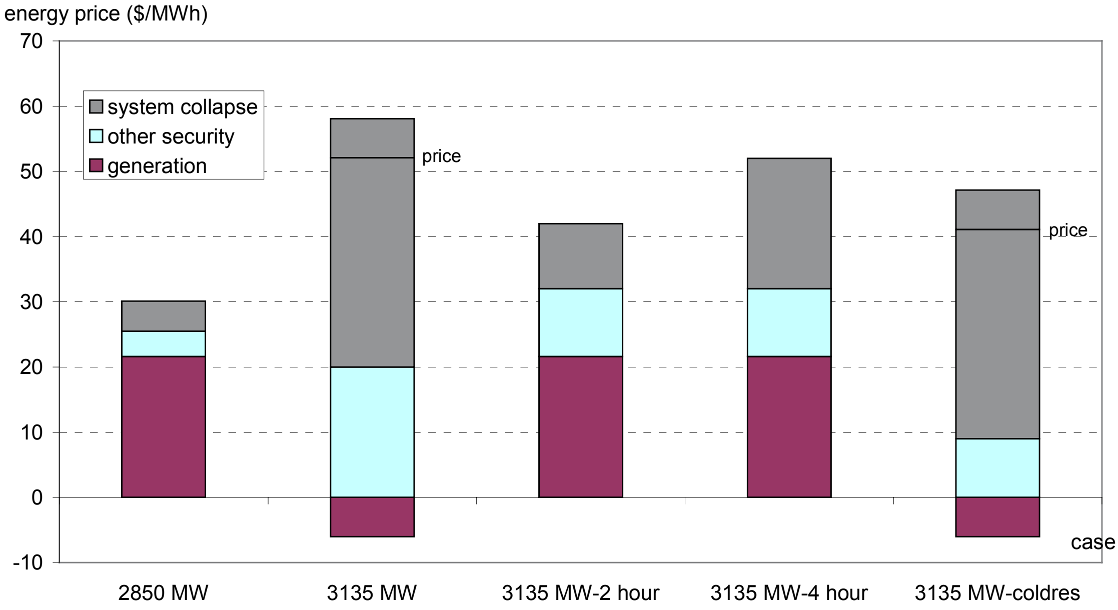

Figure 4-8: Simulated prices, including the probability of system collapse

4.4.3 Constraint on Expected Energy Not Served

The purpose of reserves is to be able to withstand the effects of load and generator (and transmission) outages with a minimum of loss of load. Instead of specifying a reserve level, a more direct way to limit the effects of outages is to constrain *EENS* directly as described in Section 4.2.9. The next table shows the results for various levels of *EENS*, including energy prices.

Table 4-9: Simulation results, D_{MAX}=3135 MW, constraint on *EENS*

EENS	energy price ($/MWh)	primary / secondary reserve (MW)	preventive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interruption cost ($)	total cost ($)
$5.2\cdot10^{-5}$	2100	170/142	42.3	0.1608 ($5.2\cdot10^{-5}$)	46151	89359	135510
$7.5\cdot10^{-5}$	2100	162/107	0.0	0.2351 ($7.5\cdot10^{-5}$)	45689	705	46394
$1.0\cdot10^{-4}$	106.5	133/100	0.0	0.3135 ($1.0\cdot10^{-4}$)	43225	941	44166
$2.5\cdot10^{-5}$	70.6	170/204	104.5	0.0758 ($2.5\cdot10^{-5}$)	45005	219673	264678

The *EENS* level in the first row is equal to *EENS* for the base case in Table 4-3, referred to the desired load level (i.e. without preventive shedding). It appears that the same level of security, as measured in 1.0-*EENS*, can be obtained with a much lower level of reserves and consequently, preventive load shedding. The *LOLP* level for the fixed reserve and *EENS*-constrained cases are 0.0145 and 0.0058 respectively. Table 4-9 shows that the constraint on *EENS* has a major impact on price.

4.4.4 Constraint on Loss Of Load Probability

As an alternative to *EENS*, the Loss Of Load Probability *LOLP* can be constrained. Results for four *LOLP* levels are shown in Table 4-10.

Table 4-10: Simulation results, D_{MAX}=3135 MW, constraint on *LOLP*

LOLP	energy price ($/MWh)	primary / secondary reserve (MW)	preventive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interruption cost ($)	total cost ($)
0.014554	37.1	50/ 79	0.0	0.7468 ($2.4\cdot10^{-4}$)	37806	2240	40046
0.003	54.7	169/88	0.0	0.2701 ($8.6\cdot10^{-5}$)	45579	811	46390
0.002	2100	165/112	174.5	0.1909 ($6.5\cdot10^{-5}$)	44383	367023	411406
0.001	2100	165/192	409.0	0.0598 ($2.2\cdot10^{-5}$)	44383[8]	859080	903463

The value 0.014554 was chosen because this is the resulting value for the 400 MW fixed reserve case in Table 4-3. It appears that this *LOLP* level can be obtained with a much lower reserve level. This results in huge savings, because in this case it is not necessary to do any preventive load shedding. In addition, generation cost is considerably lower, but *EENS* is more than four times as high. This is because the lower reserve level results in a greater impact of each outage.

[8] This value is equal to the corresponding value for the 0.002 *LOLP* level because the marginal cost of the nuclear units, whose generation is reduced in this case, is assumed to be equal to 0.

The lowest obtainable *LOLP* level without load shedding is 0.002688, caused by loss of load for the outage of one of the three largest units. To reach a lower level, it is necessary that outage of only two or one of the largest units results in loss of load. Because the maximum level or primary reserves is 170.5 MW, only two respectively one unit can generate more than 170.5 MW. This is impossible without using preventive load shedding. The results are illustrated in Figure 4-9 with a logarithmic scale for total cost. The sharp increase in total cost occurs when it is necessary to resort to preventive load shedding to obtain the desired *LOLP* level.

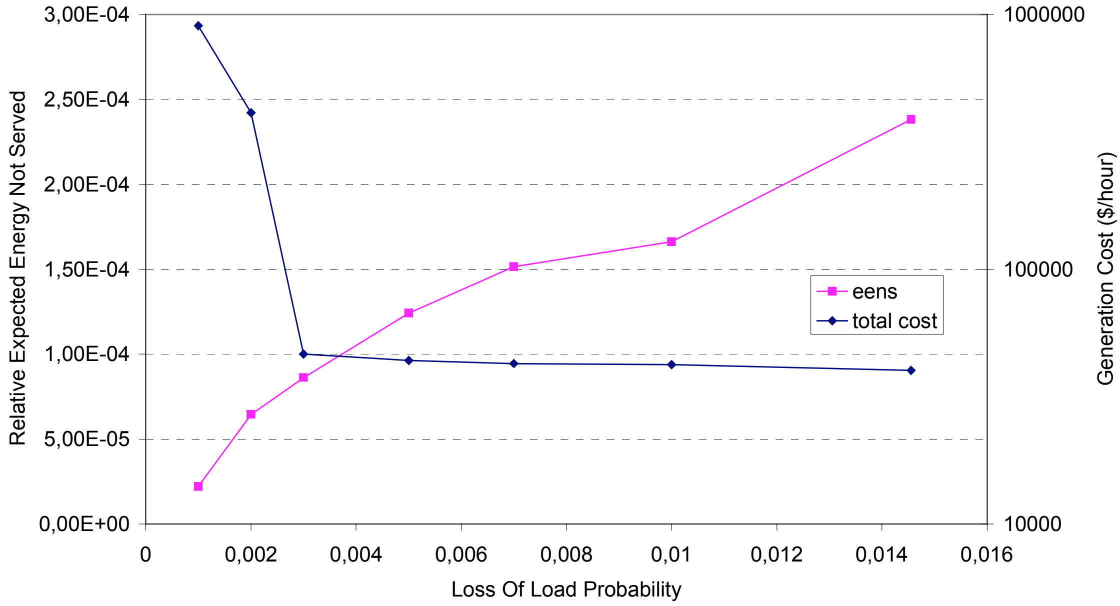

Figure 4-9: Relative *EENS* and generation cost as a function of *LOLP* constraint

4.4.5 Constraint on the Proba bility of System Collapse

A final possibility is to add constraints on the Probability of System Collapse *POSC*, as described in 4.2.8 and 4.2.9:

Table 4-11: Simulation results, constraint on *POSC*

POSC	energy price ($/MWh)	primary / secondary reserve (MW)	preventive shedding (MW)	expected energy not served (GWh / relative)	generation cost ($)	load interruption cost ($)	total cost ($)
0.000447	52.1	130/0	0.0	0.6653 ($2.1{\cdot}10^{-4}$)	39251	6200	45451
0.000377	51.9	148/2	0.0	0.5335 ($1.7{\cdot}10^{-4}$)	40452	5142	45594
0.0003	55.6	167/42	0.0	0.3295 ($6.5{\cdot}10^{-5}$)	43295	3810	47105
0.0002	2100	170/160	142.6	0.1035 ($3.5{\cdot}10^{-5}$)	48520	301634	350154

The first row corresponds to the *3135 MW* case in Table 4-8. If *POSC* is to be reduced, this can be done by including more units and moving generation from large, cheap units over to smaller and more expensive units, increasing costs. The last row gives a similar result to the base case in Table 4-3, with slightly higher preventive shedding and correspondingly reduced *EENS*. Generation costs are considerably higher, due to the redistribution of generation to reduce *POSC*.

Figure 4-10 shows *EENS* and *LOLP* as a function of *POSC*. Even for low *POSC* values, *LOLP* may be relatively high. Comparison between this and the two previous sections shows clearly that very different values result for different criteria, depending on which indicators for system security are used. A fixed reserve criterion addresses these criteria indirectly, and may turn out expensive and still not obtain the optimal results, depending on the preferred target values.

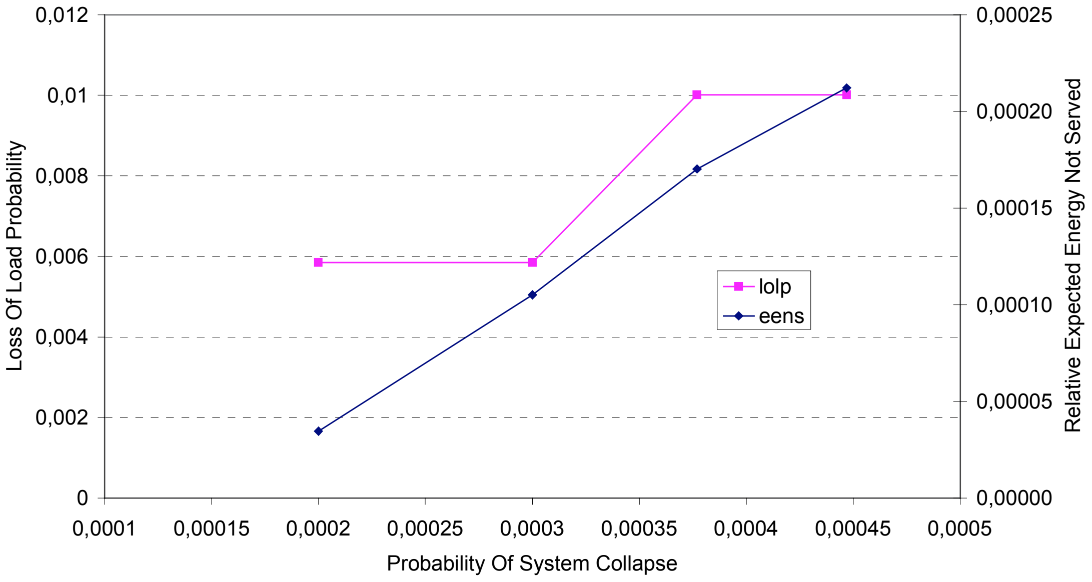

Figure 4-10: *LOLP* and *EENS* as a function of *POSC*

4.5 Conclusions regarding short-term constrained operation

In this chapter, a simulation model for short-term capacity constrained operation was developed. Simulation results under various assumptions were shown. With fixed reserve requirements to ensure system security, it can be necessary to resort to preventive load shedding to maintain these reserve levels, resulting in high social costs. Moreover, involuntary load shedding as a means to relieve capacity shortage is socially unacceptable, and will probably result in claims for market redesign.

As an alternative to the use of fixed reserves, system operators should develop a more dynamic approach to the maintenance of system security. A possible way to achieve this is by directly targeting measures of reliability like *LOLP* or *EENS*. It was shown through simulations that the same level of *LOLP* or *EENS* could be attained with lower reserve levels.

The essential instruments are typically to increase primary reserve and reduce secondary reserve, and to decrease generation on the largest units to reduce the impact of their outages. Market solutions must be developed to attain this, e.g. by establishing a market for primary reserves. It can be argued that the model is simple, but the basic ideas are valid in the real world, though more sophisticated tools need to be developed to assess the impact of outages in an operational context.

If system security is maintained by directly targeting *LOLP* or *EENS* levels, cost reductions can be obtained, and preventive load shedding can be reduced. However, the level of reliability is still exogenous. It is determined by for example the regulator or other authorities, based on assumed average values of lost load, and not by the market, based on consumers' preferences and costs. So even if prescribed reliability can be obtained more effectively, the problem of finding a market based level of reliability is not solved.

With a number of simulations on the IEEE RTS, it was again shown that the costs could be considerably reduced by utilizing demand elasticity to reduce demand. A step in the direction of market-based reliability is to let consumers contribute to the provision of reserves. A market based procedure for obtaining demand-side reserves, based on a combination of fixed annual and activation based compensation, would induce consumers to reveal their preferences with respect to reliability. The challenge is to create the necessary infrastructure to make this a feasible solution.

An important issue is the possibility of system blackout, the nightmare of every system operator and indeed, society. If the probability of blackout is too high, a particular system status may be unacceptable, although this does not show in the economic evaluations. This means that the value of loss of load used in the simulations does not reflect the excessive cost of system blackout in modern society. This can be taken into account in simulations by using extremely high costs for system blackouts, or using a constraint on the risk for such an event. The problem of estimating the risk of system collapse remains. Simulations were run with tentative system collapse probabilities, resulting in higher optimal reserve levels. In the simulations referred in this chapter, it was also shown how prices are influenced by the way the security requirement is met. If a fixed reserve requirement is used, it can be necessary to shed load to satisfy the reserve requirement, and the price will essentially equal the value of lost load, if this is done without agreements with consumers. If the security requirement is met by directly targeting the loss of load probability or expected energy not served, prices will be much lower, as long as it is possible to meet the requirements without shedding load. Thus, the level of security and the selected strategy to obtain it have a profound impact on prices.

In a restructured power system without central responsibility for system adequacy, the system operator, responsible for system security, must be prepared to face occasional capacity shortage. The key solution to this problem is to let market forces work on the demand side. To be able to operate at lower security levels with an acceptable risk of blackout, it is necessary to establish flexible load shedding routines. With these in place, it is possible to minimize the amount of load shedding during outages, and recover load as soon as generation capacity becomes available again.

4.6 References

[4.1] G.B. Sheble, G.N. Fahd, “Unit commitment literature synopsis”, *IEEE Transactions on Power Systems*, Vol. 9, No. 1, February 1994, pp. 128-135.

[4.2] Robert A. Meyer, “Monopoly Pricing and Capacity Choice Under Uncertainty”, *American Economic Review*, Vol. 65, June 1975, pp. 327-337.

[4.3] Goran Strbac, Syed Ahmed, Daniel Kirschen, Ron Allan, “A Method for Computing the Value of Corrective Security”, *IEEE Transactions on Power Systems*, Vol. 13, No. 3, August 1998.

[4.4] E.S. Huse, I. Wangensteen and H.H. Faanes, “Thermal Power Generation Scheduling by Simulated Competition”, *IEEE Transactions on Power Systems*, Vol. 14, No. 2, May 1999.

[4.5] “Applied Reliability Assessment in Electric Power Systems”, Edited by Roy Billinton, Ronald N. Allan, Luigi Salvaderi, IEEE Press, 1991

[4.6] R. John Kaye, Felix F. Wu, Pravin Varaiya, “Pricing for system security” , *IEEE Transactions on Power Systems*, Vol. 10, No. 2, May 1995.

[4.7] A. Ravindran, Don T. Philips, James J. Solberg, “Operations Research. Principles and Practice”, Second Edition, John Wiley & Sons, 1987.

[4.8] Robert Fourer, David M. Gay, Brian W. Kernighan, “AMPL. A Modeling Language For Mathematical Programming”, Boyd & Fraser Publishing Company, 1993.

[4.9] ILOG CPLEX 6.5, User’s Manual and Reference Manual, ILOG, March 1999.

[4.10] Reliability Test System Task Force of the IEEE Subcommittee on the Application of Probability Methods, “IEEE Reliability Test System” , *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-98, No. 6, Nov/Dec 1979, pp. 2047-2054

[4.11] Reliability Test System Task Force of the Application of Probability Methods Subcommittee, “The IEEE Reliability Test System – 1996”, *IEEE Transactions on Power Systems*, Vol. 14, No. 3, August 1999, pp. 1010-1020.

[4.12] Roy Billinton, Ronald N. Allan, “Reliability Evaluation of Power Systems”, Pitman Publishing, 1984

[4.13] L.T. Anstine, R.E. Burke, J.E. Casey, R. Holgate, R.S. John, H.G. Stewart, “Application of Probability methods to the Determination of Spinning Reserve Requirements for the Pennsylvania-New Jersey-Maryland Interconnection”, *AIEE Transactions on Power Apparatus and Systems*, Vol. 82, October 1963, pp. 726-735

APPENDIX A.4: THE IEEE RELIABILITY TEST SYSTEM

A4.1 System data

Computations in Chapter 5 have been done on the IEEE Reliability Test System (IEEE RTS), documented in [4.10] and [4.11]. Most cases use a single area model, represented by the IEEE RTS-79. For data not given in RTS-79, values from RTS-96 were used. The data used in this thesis are referenced in the tables below.

Table A4-1: IEEE-RTS basic data

Unit Group	Number of units	Unit Numbers	Unit Type	Max output (MW)	Min output (MW)	MTTF (Hour)	Max prim. reserve (MW)	Max sec. reserve (MW)
U12	5	1-5	Oil/Steam	12	2	2940	0	2
U20	4	6-9	Oil/CT	20	16	450	1	4
U50	6	10-15	Hydro	50	10	1980	2	10
U76	4	16-19	Coal/Steam	76	15	1960	3	15
U100	3	20-22	Oil/Steam	100	25	1200	5	20
U155	4	23-26	Coal/Steam	155	54	960	7	31
U197	3	27-29	Oil/Steam	197	69	950	9	39
U350	1	30	Coal/Steam	350	140	1150	17	70
U400	2	31-32	Nuclear	400	100	1100	20	80

Table A4-2: IEEE-RTS cost- and dynamic data

Unit Group	A0 ($)	$A1_{en}$ (c/kWh)	$A1_{prim}$ (c/kW)	SCOST ($/start)	RR (%/ min)	TSTART (min)	INIT
U12	26.00	2.63	1.0	525	8.3	5	0
U20	169.89	5.00	1.0	73	15.0	10	0
U50	0.00	0.00	0.1	0	100.0	10	1
U76	111.02	1.36	1.0	2378	2.6	10	1
U100	223.84	2.16	1.0	1585	7.0	10	0
U155	168.17	1.11	1.0	846	1.9	20	1
U197	273.31	2.16	1.0	2802	1.5	30	1
U350	190.38	1.15	1.0	6460	1.1	60	1
U400	0.00	0.00	1.0	n.a.	5.0	60	1

The cost parameters *A0* and $A1_{en}$ were calculated on the basis of the net plant heat rate in BTU/kWh and fuel prices in cent/gallon (36 and 56 for #6 and #2 oil)[9] and 26.16 $/short ton coal[10]. The piecewise linear heat rate was approximated by a linear function and converted to c/kWh. $A1_{prim}$ was chosen such that resulting cost of primary reserve was approximately 2-3 % of total running cost. Start cost was based on given data for hot start in MBTU and converted to $. Start-up time was chosen tentatively.

In some calculations, the possibility of system collapse was taken into account, given by a tentative probability for system collapse given a unit outage. For these cases, the following values were used:

Table A4-3: Tentative probabilities of system collapse

Unit Group	Pr(collapse) (%)
U12	0
U20	0
U50	2
U76	3
U100	5
U155	5
U197	8
U350	15
U400	20

A4.2 Computation of reserve requirements

A reference reserve requirement has to be established for IEEE-RTS. For this purpose, the PJM-method [4.12], [4.13] was used. The method uses the same principles as in traditional adequacy calculations, but instead of the units' annual forced outage rate FOR, their Outage Replacement Rate ORR for the period considered (e.g. 1 hour) is used. Given a two-state generator model, the probability of finding a unit on outage at time T, given that it was operating at t=0, is:

$$P(down) = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)T} \qquad \textbf{(A4-1)}$$

where λ and μ are the unit's expected failure and repair rates.

[9] September 1999 values according to US Energy Information Administration

[10] US average of coal delivered to electric utilities in 1997, source: EnergyOnLine, Electricity Statistics.

For the short period considered, the repair time is neglected (μ=0), resulting in:

$$P(down) = 1 - e^{-\lambda T} \approx \lambda T = \frac{T}{MTTF} = ORR \qquad \textbf{(A4-2)}$$

where MTTF is the unit's mean time to failure. In the computations, a period of one hour is used, resulting in ORR=1/MTTF.

The system's individual outage probabilities were calculated using the following simple algorithm:

$$\begin{aligned} &Pr(0) = 1 - \prod_{i \in G}(1 - ORR_i) \\ &Pr(\sum_{i \in O} c_i) = \prod_{i \in O} ORR_i \cdot \prod_{i \notin O}(1 - ORR_i) \end{aligned} \qquad \textbf{(A4-3)}$$

where Pr(.) are the individual outage probabilities, G the set of all generators, O the set of outaged generators and c_i the capacity of unit i. Single, double and triple outages are considered. After computing all probabilities, the results were sorted, truncated and accumulated. This method gives the exact probabilities, and is feasible for small systems. Otherwise, the recursive algorithm referred in [4.12] must be used with discrete outage levels and steps of for example 20 MW.

The results for IEEE-RTS are shown in Figure A4-11. With a requirement of probability of loss of load of less than 0.001, a reserve level of 400 MW is required if there is no load uncertainty. Using the technique described in [4.13], reserve requirements with load uncertainty are given in Table A4-4.

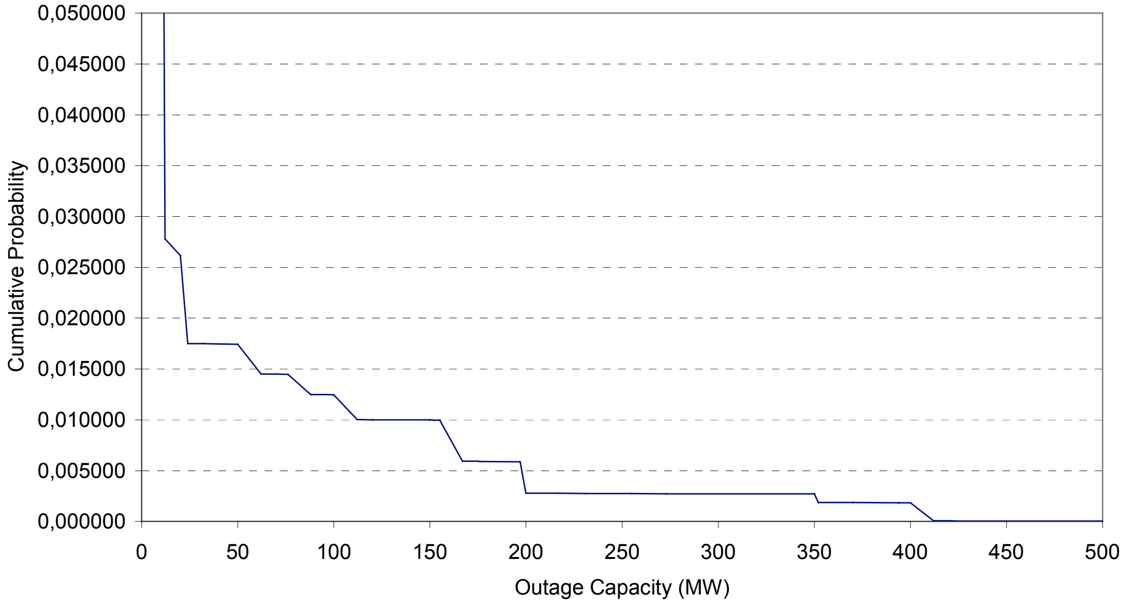

Figure A4-11: Cumulative outage probability based on ORR for IEEE RTS.

Table A4-4: Reserve requirements to obtain a cumulative outage probability <0.001

load uncertainty (standard deviation in %)	required reserve level
0	402
2	402
4	438
6	525
8	605
10	697

Chapter 5: INTEGRATED PRICING OF ENERGY AND ANCILLARY SERVICES

The most important special property of a power market is the necessity to maintain a continuous balance between supply and demand, caused by the impossibility to store electricity economically. Because of this, it is necessary to maintain certain levels of reserves that differ with respect to their time and frequency responses. In the previous chapter, a distinction was made between primary and secondary reserves and their use in a post-contingency situation. Focus was on optimal provision of necessary reserves, moving to optimal provision of reliability, characterized by various criteria. However, much research will be necessary before the principles in the previous chapter can be readily applied in existing large-scale power systems. In the near future, reliability will in most cases be taken care of by defining required levels of reserves.

In restructured systems, reserves are explicitly provided through ancillary services[1]. These can be defined as "activities that pertain to the provision of all electric services necessary for efficient and reliable generation, transmission and delivery of active power with a sufficiently stable frequency and voltage" [5.1]. The availability of sufficient generation capacity is essential with respect to reliable generation and frequency issues. Thus, there is a link between generation capacity on the one side, and the provision of ancillary services on the other.

In this context, a reasonable hypothesis is that the opportunity to earn revenues in ancillary services markets would increase the profitability of generation investments, and thus reduce the peaking capacity problem. To investigate this possibility, a model for the simulation of a market for energy and ancillary services is used in this chapter.

In Section 5.1, an integrated market model for energy and ancillary services is described. In Section 5.2 the mathematical formulation of the model is given, while Sections 5.3 and 5.4 describe the solution procedure. Sections 5.5 and 5.6 describe the results from simulations on a test system, while Section 5.7 draws some conclusions.

5.1 An integrated market for energy and ancillary services

5.1.1 Existing solutions and the proposed model

In the Nordic market (e.g. [5.2]), hourly energy prices for a 24-hour period are determined by the Power Exchange (PX), based on price versus quantity bids for supply and demand. National System Operators operate separate regulating markets to acquire secondary reserves.

[1] In traditional systems, these services were provided more implicitly through the constraints imposed on the Unit Commitment and dispatch solutions.

As far as the provision of other ancillary services is compensated, it is based on negotiated terms in the Norwegian market

The Californian solution for the 24-hour day-ahead market is similar to the Nordic. In addition, short-term markets for four ancillary services are recognized: regulation reserve, spinning reserve, non-spinning reserve and replacement reserve [5.3]. Separate markets exist for each of these.

In contrast, the UK Pool[2] performs a unit commitment (UC) and dispatch based on bids from all generators. The bids consist of a number of parameters for each unit, which serve as a proxy for the real data. Ancillary services are compensated, and partly provided through explicit markets.

Numerous papers have proposed various market structures. Some examples are given: Hao et al. [5.6] present a procedure for calculation of customer payment minimizing prices for energy. Aganagic et al. [5.7] include spot pricing of reserve generation and transmission capacities as well as demand side bidding. Richter and Sheblé [5.8] propose a market structure with ENSERVCO's and ANSILCO's, providing energy and ancillary services respectively. Hirst and Kirby [5.9] use a semi-static approach in a comprehensive study of prices for ancillary services. Bakirtzis describes a model for joint energy and reserve dispatch [5.10].

5.1.2 Mutual exclusivity

The examples given, and not at in the least the market solution in California, illustrate a fundamental problem confronting the designer of short-term markets for ancillary services: energy and (reserve type) ancillary services are mutually exclusive products. This means that a generator has to choose to which degree it wants to produce energy or one or more reserve services – it cannot produce (the maximum of) both simultaneously. For example, a generator with a capacity of 100 MW cannot at the same time produce 100 MW of electricity and 20 MW of reserves, because the production of reserves implies the ability to increase the generation of electricity. This mutual exclusivity poses substantial challenges both for the formulation of bids and for clearing the market. In California, markets are cleared sequentially in a quality-of-service order: regulation first, followed by spin, non-spin and finally, replacement [5.5]. In this sequential market-clearing process, capacity that is won in a previous auction is subtracted out from the capacity that is bid into the subsequent markets. For example, if a participant bids 100 MW of a generating unit into regulation and 200 MW into spin and wins 50 MW in regulation at the regulation bid price, then 200 MW – 50 MW = 150 MW is actually bid into the spin market at the spin bid price. If 80 MW is won in the regulation market, then only 120 MW is actually bid into the spin market at the spin price. If 100 MW is bid into regulation and 50 MW is bid into spin, even if *none* of the 100 MW of bid into regulation is taken in the regulation market, only 50 MW is bid into the spin market.

[2] The Pool is scheduled to be replaced with a new market mechanism in the autumn of 2000, cf. Section 2.5.

The fact that energy and the various reserve services are mutually exclusive from the generator's point of view is an argument for an integrated clearing of these markets, cf. also [5.11]. Here a market structure somewhat between the UK and Californian solutions is proposed. It is based on central dispatch by an independent Market Operator (MO)[3], but with explicit short-term pricing of ancillary services. The solution has the following motivation:

- It provides a way to estimate prices for ancillary services, necessary to asses the question posed in the introduction to this chapter.
- It represents an alternative market design. No final "best" solution has emerged so far, and the study of alternatives provides new insights and may suggest better market designs. However, present developments seem to be away from UK pool type solutions.
- Computed prices for an integrated ideal market for energy and ancillary services serve as benchmarks for other market solutions.
- It may serve as a tool for multi-plant owners in a market like the Californian, to optimize their own production of energy and ancillary services.

5.1.3 Ancillary services

The following ancillary services are defined:

- Primary control to provide primary reserve, with a response time of seconds. This is used to control short-term load fluctuations and to some degree uncertainty, and naturally provided by spinning units.
- Secondary control, with a response time varying from seconds to minutes, and full activation within 10-15 minutes. This is used in the first place to cope with outages in generation or transmission but also for greater deviations from the load forecast. It is often but not everywhere provided by spinning units.
- Tertiary control, which has a response time of 30-60 minutes, and is used to restore secondary reserve. Depending on system characteristics and requirements, it is provided by spinning or non-spinning units.

Other types of ancillary services can be readily defined, provided they can be characterized by response time and spinning requirement.

5.1.4 Market operation

The following assumptions are made:

[3] In real systems there is often a division of tasks between a Market Operator and (Independent) System Operator or (I)SO. In the current context, this distinction is not relevant, and the market is assumed operated by one entity, called the Market Operator or MO.

- There are a number of agents, each owning exactly one generator or representing a demand segment. This represents perfect competition, where no market participants own a significant share of generation, and all are price takers.
- Consumers pay separately for energy on the one hand and transmission/distribution on the other hand, as is becoming common in deregulated power markets. In the demand function for energy, the effect of the transmission/distribution tariffs is taken into account.
- Energy is purchased by many consumers, while ancillary services are purchased as a public good by the MO only. The costs are compensated e.g. as part of the transmission/distribution tariffs.
- The transmission grid is ignored in the calculation of the market equilibrium. To what degree this is a realistic assumption depends on the market structure. If the market uses a principle of one "system price", the MO buys power on the high-price side and sells on the low-price side of a transmission constraint, and the major part of the market does not see the constraint. In this case the assumption is realistic.
- As a consequence of the previous assumption, voltage control / reactive power is assumed to be acquired on other, presumably long-term markets.
- The necessary level of ancillary services is assumed exogenous. Theoretically, the optimal level would be obtained where the marginal cost of an ancillary service is equal to its marginal benefit, given by the expected marginal reduction in outage costs, cf. Chapter 4. In practice, this expected marginal reduction is difficult to calculate, and standards based on traditional criteria are agreed and imposed by e.g. the MO.
- Ancillary services may be provided by consumers. Traditionally reserves were always provided by generators, but in restructured market it is plausible to assume that some consumers may be willing to reduce demand on short notice, if they are given satisfactory compensation. For example warm water boilers with sufficient volume may be switched off during a few critical peak hours without loss of comfort. In this context, demand reduction is equivalent to an increase in generation.

A pool-based market structure is envisaged, which is cleared sequentially, hour for hour. Starting with the first hour, all agents prepare their bids. Intertemporal dependencies are considered by each individual agent, using a price forecast for the remaining price period, in the same way as in [5.13].

In this market structure, the MO receives all the bids for the current hour, and computes the market solution: prices and volumes for each agent for energy and the three ancillary services. From this market solution, each agent knows his commitment state, and can proceed with preparing his bid for the next hour.

In this way, all hours for one twenty-four hour period are cleared. It is appropriate to question the quality of the solution in terms of optimality: the introduction of competition ideally should not affect operating efficiency adversely. It is obvious that the agents' price forecasts play an essential role. If the forecast was close to the final prices, the agents' commitment decisions will have been taken on the correct basis, and they will be optimal *ex post*. If, however, the forecast was wrong, the decisions would be sub-optimal *ex post*, and the market solution will not be cost-minimizing. To approach the cost-minimizing solution, it may be necessary to allow for several iterations in the whole process. The results from a previous iteration may then be used by the agents as a price forecast. Naturally, the number of iterations will be limited if used in a real market. As agents gain experience with the market, their initial price forecasts may normally be good enough.

5.1.5 Market prices for energy and ancillary services

Prices and volumes are calculated in a centrally run optimization process as described in the previous section. The parameters for the optimization are given by way of the market participants' bids.

Prices should be such that volumes and prices are consistent with market participants' self-dispatch decisions – i.e. given these prices, market participants would have chosen the same production of energy and ancillary services as the outcome of the optimization program.

The basic challenge occurs because of startup costs and the operation-dependent constant term in the cost function. Because of these, the marginal cost of the marginal (price-setting) unit is often lower than the average cost of this unit. To make production profitable for this unit, it is necessary to raise the price above its marginal cost. However, if this is done, other units will increase their production, and total generation will exceed demand. Thus in many cases it is not possible to find a price that is such that generation equals demand. To solve this problem, an additional "balancing" product is introduced, defined as the difference between total generation and demand. This is illustrated in Figure 5-1. The left panel shows the classical solution, where price $\lambda'_{en,t}$ equals marginal cost for generator i. The generator has a minimum generation constraint of $PMIN_{en,it}$. Because of the startup and operation-dependent constant terms, the cost of generating $PMIN_{en,it}$ is considerably higher than the marginal-cost curve would suggest, and is represented by the hatched rectangle. With a price $\lambda'_{en,t}$ and a production $p_{en,it}$ the part of this rectangle exceeding $\lambda'_{en,t}$ is greater than the light shaded triangle. Consequently the generator would decide not to produce at all.

The right hand side of Figure 5-1. shows the situation where the price is $\lambda_{en,t}$ and generator i would choose to produce $p_{x,it}$. However, it may happen that total generation exceeds demand in this case because the price had to be increased to commit sufficiently many generators, cf. Section 5.3.2. From this situation, all generators are given the opportunity to reduce their output by balancing. After balancing, generator i produces $p_{en,it}$ for which it receives the price $\lambda_{en,t}$. Generator i does not increase generation, because it also "produces" a balance production (in reality a reduction in output) of s_{it} at a price ν_t. The combined generation of $p_{en,it}$ and s_{it} constitutes the optimal solution for each unit.

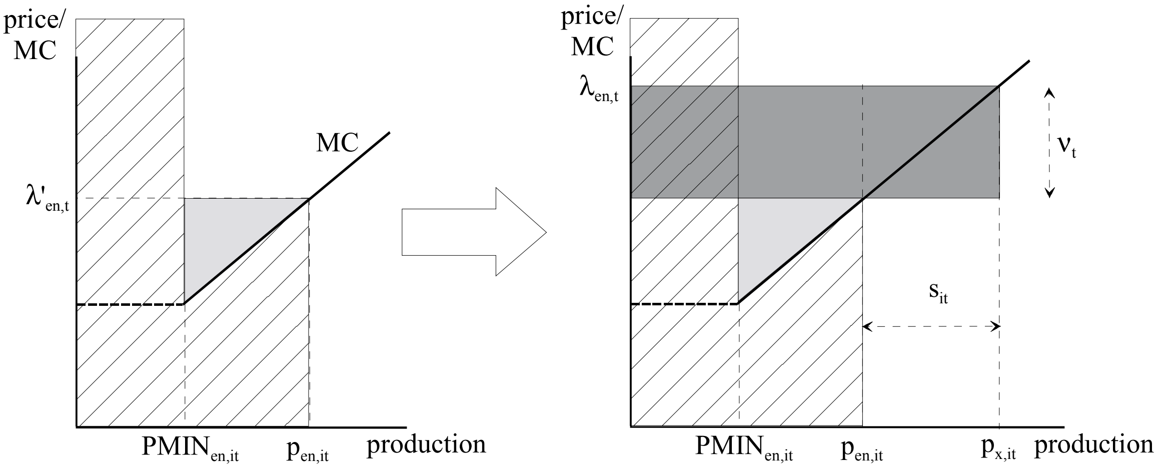

Figure 5-1: Marginal cost, balancing production s_{it} and its price ν_t.

At the equilibrium, total production costs are given by the hatched area. Revenues from energy sales are given by the product $p_{en,it} \cdot \lambda_{en,t}$, and from balancing by $s_{it} \cdot \nu_t$. The additional income in the right hand panel compared with the left hand panel is represented by the dark shaded area. In this case revenues exceed costs, and the generator will produce at the equilibrium point $p_{en,it}$, where marginal costs equals the difference $\lambda_{en,t}$ - ν_t. It can be seen that this is optimal for the generator with given prices $\lambda_{en,t}$ and ν_t because an increase dP in generation would change revenues with $dP \cdot (\lambda_{en,t} - \nu_t - MC^+)$, where MC^+ is slightly higher than MC because of the increase in production. This is negative, because in the optimum $MC = \lambda_{en,t} - \nu_t$. Similarly a decrease dP in generation would change revenues with $dP \cdot (-\lambda_{en,t} + \nu_t + MC^-)$, where MC^- is slightly less than MC, and again this is negative. The size of the additional revenue is greatly exaggerated in Figure 5-1 for clarity purposes. In the example in Section 5.5.2, the additional revenue is 9 % of total generator revenues in the hour with highest prices, on average 1 % for all hours. These costs are assumed covered as a part of the transmission tariffs.

5.2 Mathematical formulati on

5.2.1 Symbol definitions

In principle, the same symbols are used as in the model described in the previous chapter. However, the time horizon of the present problem is different, so it must be kept in mind that the time variable *t* represents different units in the respective models (typically, it will be one hour in the present model). Furthermore,

Sets:

P - the set of commodities: {“en”,”prim”,”sec”,”tert”}, where “en” is an abbreviation for energy (MWh/hour), “prim” for primary control, “sec” for secondary control and “tert” for tertiary control (all MW)

Indexes:

t - time period, $t \in \{1,\dots,T\}$

i - generation unit (agent), $i \in \{1, \dots, I\}$

j - demand segment (agent), $j \in \{1, \dots, J\}$

k - commodity, $k \in P$

Constants:

T - number of time periods

I - number of generation units (suppliers)

J - number of demand segments (consumers)

$PMAX_{ki}$ - maximum value of commodity k produced by unit i if running

$PMIN_{ki}$ - minimum value of commodity k produced by unit i if running

$PMAX0_{ki}$- maximum value of commodity k produced by unit i if not running

CL_i - additional startup cost for unit i after infinite down time (cf. equation **(5-1)**)

CH_i - hot startup cost for unit i (i.e. startup cost after a theoretical zero down time)

TC_i - cooling time constant unit i

$A0_{it}$ - constant term in running-cost function of unit i, period t

$A1_{kit}$ - coefficient of linear term in cost function of unit i, period t

$A2_{it}$ - coefficient of quadratic term in running-cost function of unit i, period t

TUP_i - minimum up time unit i after start

$TDOWN_i$- minimum down time unit i after stop

R_{kt} - Requirement (demand) for commodity k in period t

$TLEN_t$ - Length of time period t in basic time units

$IDIFF_{it}$ - Future benefit for unit i of running in period t compared with not running in period t

$DMAX_{jt}$ - maximum demand of segment j in period t

$DMIN_{jt}$ - minimum demand of segment j in period t

The index t for the constants $A0$, $A1$, $A2$ indicates their dependence on the length of the time step.

Variables:

u_{it} - commitment status of unit i in period t, $u_{it} \in \{0,1\}$. 1 indicates the unit is running

p_{kit} - production of commodity k by unit i in period t

$\underline{p}_{it}$ - vector of all p_{kit}, $k \in P$

s_{it} - “balancing production” of unit i in period t

$c_{it}(.)$	-	running cost of unit *i* in time period *t*
d_{jt}	-	customer *j*'s demand for energy in period *t*
$toff_{it}$	-	number of basic time units unit *i* has been off at the start of period *t*
ton_{it}	-	number of basic time units unit *i* has been on at the start of period *t*
λ_{kt}	-	price for commodity *k* in period *t* (in the market clearing problem)
$\underline{\lambda}_t$	-	vector of all λ_{kt}, $k \in P$

The basic time unit may be chosen arbitrarily. A natural choice is the length of the shortest time step used.

5.2.2 Generator model

A quadratic cost function is assumed for the generators. A cost is accrued for the provision of ancillary services. For example as pointed out in [5.14], the provision of frequency regulation leads to reduced fuel efficiency. This has, somewhat simplified, been modelled as a constant specific cost. In general, the cost will be zero for ancillary services with longer time frames. The maximum values for reserve contributions by a single generator $PMAX_{ki,}$ is limited by constraints on its ramping rates and, in the case of primary reserve, also the unit droop and the frequency deviation. Thus the model equations for generator *i* are:

$$\begin{aligned} c_{it} &= A0_{it}u_{it} + \sum_{k=1}^{K} A1_{kit}\,p_{kit} + A2_{it}\,p^2_{en,it} \\ &+ \left(CH_i + CL_i\left(1-e^{-toff_{it}/TC_i}\right)\right)\cdot\left(1-u_{i(t-1)}\right)\cdot u_{it} \end{aligned} \quad \textbf{(5-1)}$$

$$\begin{aligned} p_{kit} \le\ & PMAX_{ki}\cdot u_{it} + \\ & PMAX0_{ki}\cdot(1-u_{it}) \qquad \forall k \end{aligned} \quad \textbf{(5-2)}$$

$$\sum_{k\in P} p_{kit} \le PMAX_{en,i} \quad \textbf{(5-3)}$$

$$p_{en,it} - p_{prim,it} - PMIN_{en,i}\cdot u_{it} \ge 0 \quad \textbf{(5-4)}$$

$$ton_{it} = \left(ton_{i(t-1)} + TLEN_{t-1}\right)\cdot u_{i(t-1)} \quad \textbf{(5-5)}$$

$$toff_{it} = \left(toff_{i(t-1)} + TLEN_{t-1}\right)\cdot(1-u_{i(t-1)}) \quad \textbf{(5-6)}$$

$$u_{it} \ge u_{i(t-1)} - ton_{it}/TUP_i \quad \textbf{(5-7)}$$

$$u_{it} \le u_{i(t-1)} + toff_{it}/TDOWN_i \quad \textbf{(5-8)}$$

$$p_{kit} \ge 0 \qquad \forall k \quad \textbf{(5-9)}$$

$$u_{it} \in \{0,1\} \quad \textbf{(5-10)}$$

The obvious $\forall t$ has been omitted in these equations. Fuel and startup costs are defined in **(5-1)** in the usual way and the cost for ancillary services are added. Equation **(5-2)** expresses that the maximum values of energy and the various ancillary services depend on the commitment state of the unit. Naturally $PMAX0_{en,it}$ and $PMAX0_{prim,it}$ are 0, while the values for the other reserves may be greater than 0, if the particular ancillary service does not require the providing units to be spinning. **(5-3)** expresses that the sum of active generation and ancillary services never can exceed the generator's capacity. **(5-4)** avoids the provision of primary reserve by units running at minimum power. Alternatively, separate products could be defined for downward and upward regulation, as has been done in California. In this case, **(5-4)** would apply to downward regulation only, while this would not be included in the sum of **(5-3)**. **(5-5)**-**(5-8)** define up- and downtimes, and restrict these to their minimum values, while **(5-9)** is the general requirement that generation and reserves are positive.

5.2.3 Demand model

Traditionally demand has been considered inelastic in the short run. This is still a reasonable assumption for a considerable part of demand, but in deregulated markets an increasing part of demand is becoming elastic. In view of the other parts of this thesis, it is natural to investigate the effect of flexible demand on the ancillary services markets. Therefore, the same model as in Chapter 4.2 is used.

5.2.4 Bid formulation

Given a price forecast, the agent's problem is to prepare a bid for one hour and for four commodities: energy and three ancillary services. The commodities are strongly inter-related, and in addition, inter-temporal constraints must be taken into account. This suggests that it will be difficult to prepare explicit price-quantity bids that consider the mutual dependencies. Thus, the bid is formulated as an optimization problem, and the task of the MO will be to solve this problem on behalf of all agents.

In order to construct his bid for hour t, each agent has to calculate the future consequences of being committed in period t. This decoupling in time is based on the model described in [5.13], and shortly outlined here. A traditional UC dynamic programming (DP) procedure is used, with the objective:

$$\underset{\underline{p}_{i\tau},u_{i\tau}}{\mathrm{MAX}} \quad w(u_{it}) = \sum_{\tau=t+1}^{T} \left[\sum_{k\in P} [\lambda_{k\tau} p_{ki\tau}] - c_{i\tau}(u_{i\tau}, u_{i(\tau-1)}, toff_{i\tau}, \underline{p}_{i\tau}) \right] \quad \textbf{(5-11)}$$

subject to **(5-1) - (5-9)**. In this formulation, only compensation for having reserves available is assumed. Strictly, a term representing compensation for activated reserves should be added. If the probability of activating the reserve is small, and the costs are moderate, the contribution from this term may be neglected in the optimization problem.

$\lambda_{k\tau}$ represent the price forecasts for energy and ancillary services in period τ. **(5-11)** is solved using a backward DP procedure, where the optimal production in each period is solved with a quadratic programming routine.

By solving **(5-11)** for u_{it}=0 and u_{it}=1 respectively, the agent can calculate the future benefit of his commitment decision in period t, $IDIFF_{it} = w(1) - w(0)$ and prepare his bid for this period. The bid is the solution of the maximization of the agent's profit function:

$$\underset{\underline{p}_{it}, u_{it}}{\text{MAX}} \quad z(\underline{\lambda}_t) = \sum_{k \in P} [\lambda_{kt} p_{kit} \ ] - c_{it}(\cdot) + IDIFF_{it} \cdot u_{it} \tag{5-12}$$

subject to **(5-1)** - **(5-9)**.

The solution of **(5-12)** is a multi-valued price-volume relation from җ^{K}→җ^{K}, where K is the number of commodities. The bid is submitted to the MO as a statement of the optimization problem with its constraints. This form of bid will be too complicated for many market agents, but simpler forms of bids, like the usual price-volume relations, may be regarded as subsets of this more general form.

5.3 Single hour market clearing

The MO receives bids for energy and ancillary services from all market agents in the general form of **(5-12)** with its constraints. Because demand is inelastic (cf. the modelling of elastic demand in the previous chapter), the profit maximization problem to solve for the MO can be converted into the cost minimization problem **(5-13)**:

$$\underset{\underline{p}_{it}, u_{it}}{\text{MIN}} \left[\sum_{i=1}^{I'} c_{it}(\cdot) - IDIFF_{it} \cdot u_{it} \right] \tag{5-13}$$

subject to **(5-1)** - **(5-9)** $\forall i$ and:

$$\sum_{j=1}^{J} DMAX_{jt} - \sum_{i=1}^{I'} p_{en,it} = 0 \tag{5-14}$$

$$\sum_{j=1}^{I'} p_{kit} \geq R_{k,t} \ \forall k \neq "en" \tag{5-15}$$

and where I' equals I increased with the number of "generators" representing elastic demand.

This formulation of **(5-15)** assumes that each form of reserve only can be used for its own purpose. However, as has been discussed in California and New England, higher quality services can be used to satisfy lower quality requirements, e.g. primary reserve may be used as secondary reserve etc. This "rational buyer" concept (or "cascading markets") can be modelled by a reformulation of **(5-15)**:

$$\sum_{i=1}^{I'} \left[p_{prim,it} \right] \geq R_{prim,t}$$

$$\sum_{i=1}^{I'} \left[p_{prim,it} + p_{sec,it} \right] \geq R_{prim,t} + R_{sec,it}$$

$$\sum_{i=1}^{I'} \left[p_{prim,it} + p_{sec,it} + p_{tert,it} \right] \geq R_{prim,t} + R_{sec,it} + R_{tert,it}$$

5.3.1 Branch and Bound (BB) solution

A branch and bound procedure can solve **(5-13)** - **(5-15)** with constraints **(5-1)** - **(5-9)** directly. A general-purpose QP solver was used to solve the quadratic problem, while a dedicated branch and bound procedure was user written. In this case, the dual values from the constraints **(5-14)** and **(5-15)** are interpreted as prices. However, with these prices generators do not necessarily cover their costs, because neither startup costs nor the constant terms $A0_{it}$ are taken into account in the dual values, cf. Willams [5.16]. If the problem is not too large, it is possible to find the optimal production schedule, but feasible market prices are not obtained by this procedure.

5.3.2 Lagrange Relaxation (LR) solution

To be feasible in a market environment, the solution to **(5-13)** - **(5-15)** with constraints **(5-1)** - **(5-9)** should give prices that satisfy each market agent. A feasible market solution implies prices that would yield the same outcome if each agent would self-dispatch, given these prices and his expectations of prices for future periods. It is easy to find simple examples where this is not possible (also cf. [5.15]), so two additional assumptions must be made:

1. When agents are economically indifferent with respect to certain production levels, it is accepted that the MO decides on the final levels[4].
2. An additional balancing production is introduced as a means to obtain equality between total generation and demand (cf. Section 5.1.5).

The problem defined by **(5-13)** - **(5-15)** bears strong resemblance to the traditional UC problem. This suggests the use of Lagrange Relaxation. The dual LR formulation is:

[4] This is not a special requirement for the market concept described here. As shown by Madrigal and Quintana in [5.17], there are possibilities for multiple optimal solutions even in simple uniform first-price auctions.

$$\underset{\underline{\lambda}_t}{\text{MAX}} \quad \underset{\underline{p}_{it}, u_{it}}{\text{MIN}} \quad \sum_{i=1}^{I'} \left[c_{it}(\cdot) - IDIFF_{it} \cdot u_{it} - \sum_{k \in P} \lambda_{kt} p_{kit} \right] \tag{5-16}$$

subject to **(5-1)** - **(5-9)** $\forall i$, where **(5-14)**-**(5-15)** have been relaxed.

The solution of **(5-16)** is obtained, basically using a subgradient procedure. As pointed out in the literature, it is generally hard to find good primal feasible solutions from the optimal dual solution. Procedures where primal feasible solutions are generated during the iteration procedure are used in most references (cf. [5.18], [5.19]). On this background, the following algorithm was designed:

1. Set initial price forecast λ_{kt} $\forall k,t$.
2. Solve **(5-11)** twice for each agent with his current price forecast to find the $IDIFF_{it}$'s.
3. Solve **(5-12)** for all agents. Verify satisfaction of the global constraints **(5-14)** - **(5-15)**. If total generation is greater than or equal to demand and reserve requirements can be satisfied (if necessary by rescheduling of committed units), go to 4. If total generation is less than demand, go to 5.
4. There is a generation surplus with given prices. Solve the primal problem **(5-13)** with given u_{it}, resulting in new multipliers λ_{kt} as the dual values of **(5-14)**-**(5-15)**. Update the multipliers according to:

   $$\lambda_{kt}^{next} = \alpha \cdot \lambda_{kt} + (1-\alpha)\lambda_{kt}^{old} \tag{5-17}$$

   where $0 \le \alpha \le 1$. Go to 6.
5. Not enough units are committed to satisfy the power balance. Use subgradient-updating:

   $$\lambda_{kt}^{next} = \lambda_{kt}^{old} + \frac{1}{a + b \cdot ITER} \cdot \frac{R_{kt} - \sum_i p_{kit}}{R_{kt}} \tag{5-18}$$

   where ITER is the iteration number.
6. If improvement in the primal objective has occurred in the last N iterations, go to 2. Otherwise, finish.

5.4 Multiple bidding rounds

The process described in the previous sections describes the market clearing process within a single hour. To be able to construct his bid for one hour, the generator must:

- know it's commitment status at the previous hour
- use a price forecast for a number of future hours

Because of the first requirement, a sequential bidding process is proposed, where one hour is cleared before the bidding process for the next is started. If the original price forecasts are reasonably correct, the resulting commitment and dispatch will come close to the least cost solution. However, if the forecast deviates substantially from the final prices, the solution will be suboptimal. A bidding process with multiple bidding rounds can be simulated by employing an outer loop upon the market clearing process described in 5.3, where prices from the last round are used to adjust the forecast used in the next round. An obvious adjustment is simple linear weighting of the price forecast with the prices from the last round, analogous to **(5-17)**.

In a real market, the number of bidding rounds will naturally be limited. However, in a simulated market, it is possible to run many rounds to search for the global optimum.

5.5 Results from test calcula tions

5.5.1 Model verification

To verify the model described in this chapter, a comparison was made with the 10-unit model used in [5.12] and [5.13]. Details of the model, together with additional data used in the subsequent sections are given in Appendix A.5 at the end of this chapter. To enable a comparison of the results, the requirements for primary and tertiary reserves (cf. **(5-15)**) are set equal to zero. To make direct comparison possible, the values $PMAX_{ki}$ were set equal to $PMAX_{en,i}$ for all k. In model terms this means that **(5-2)** becomes non-binding for k Í ”en”.

In the first comparison the BB algorithm was used to solve the problem **(5-13)**-**(5-15)**. As discussed before, this does not give feasible market prices, but it guarantees the optimal (minimal cost) solution of **(5-13)**-**(5-15)**, i.e. within each hour. The resulting cost is USD 47262, better than any other solution referred in Table 6.1 in [5.12], indicating the present method is able to find good solutions to the UC problem. The resulting commitment schedule is given in Table 5-1. The differences with the results in [5.12] are underlined.

Table 5-1: Unit commitment plan test system, Branch and Bound

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Subsequently, a simulation was run where the LR algorithm described in Section 5.3 was used instead of BB. Total cost for this solution was USD 47293. The resulting commitment plan is shown in Table 5-2, which only shows the units with a different commitment than the previous solution. The differences with Table 5-1 are underlined.

Table 5-2: Unit commitment plan test system, LR algorithm (differences with Table 5-1)

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Total costs are 0.07 % higher than for the BB solution, but among the best referred in Table 6.1 in [5.12]. The next two tables show unit profits for the respective methods. The numbers in these tables do *not* include expected future income $IDIFF_{it}$ defined in Section 5.2.4.

Table 5-3: Unit profit (USD), Branch and Bound

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	15	10	12	12	11	13	17	13	11	10	12	12	10	9	9	7	6	10	8	7	6	9	13	25
2	6	-1	2	2	1	4	9	4	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	6	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	10	2	6	5	4	8	15	8	4	2	6	5	3	1	0	-2	-4	0	0	0	0	0	0	0
5	-66	11	17	15	14	19	30	19	14	11	17	15	12	10	8	5	2	12	6	4	2	10	18	47
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-147
7	20	-4	8	5	2	13	34	13	2	-5	8	4	-3	-7	-10	-17	-23	-2	-14	-18	-22	-6	10	75
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	27	14	21	19	17	23	34	23	17	14	21	19	15	13	11	8	4	15	9	7	5	13	22	55
10	55	40	48	45	43	51	64	51	43	39	48	45	40	37	35	30	25	40	32	29	26	38	49	88

Table 5-4: Unit profit (USD), LR algorithm

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	15	10	9	7	6	8	10	9	10	10	11	10	12	9	9	8	7	8	8	6	-3	2	6	55
2	6	-1	-3	-5	-6	-4	-1	-2	-1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
3	6	-2	-4	-7	-8	-6	-2	-3	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	10	2	0	-2	-4	-1	2	1	3	3	4	3	6	1	1	-1	-1	0	0	-3	-16	-9	-3	90
5	-66	11	8	5	2	6	11	10	12	12	13	13	16	10	9	7	5	9	8	5	-14	-5	3	123
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	20	-4	-10	-17	-22	-14	-3	-6	-2	-1	0	-1	6	-7	-9	-10	-11	-9	-12	-21	-60	-37	-20	349
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	27	14	11	7	5	9	15	13	15	16	16	16	19	13	12	12	12	11	11	6	-15	-4	6	226
10	55	40	35	30	26	32	40	38	40	41	42	41	45	37	36	33	30	34	31	26	-5	13	27	189

With the BB method in Table 5-3, there is no constraint on profit. The effect of this is easiest to see on Unit 6 in hour 24 – clearly the unit does not earn it's startup costs, and would not be willing to run, if it was given a choice.

In Table 5-4, profits *including expected future income* should be positive. Negative profits can occur for three reasons:

- Unit must run because of minimum up time
- Expected future surplus exceeds the accumulated deficit, e.g. Unit 9 in hours 21 and 22
- Final prices differ from expected prices

The last reason is related to the multiple bidding rounds. Expected prices in one round are based on the results from the previous two rounds. It is well known that iterative procedures like the current one tend to end up in an unstable behaviour between two or a few solutions. These solutions will often have different prices. Thus the price forecast the least cost solution is based upon may be too high, resulting in an *ex post* wrong commitment decision. This is obviously the case for Unit 7 in the example above. A further illustration of the solutions is given in Figure 5-2.

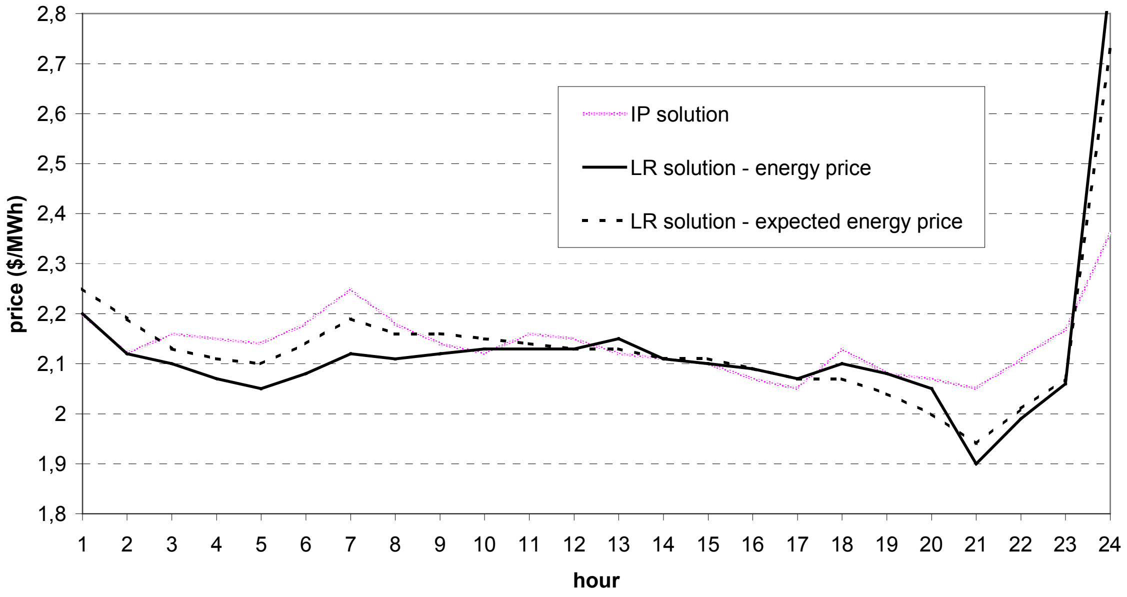

Figure 5-2: Prices and marginal costs

The grey line shows the results for the BB solution. The reason that prices are higher than for the LR solution in most hours is that the BB process finds cheaper solutions with fewer units committed, resulting in higher marginal costs. The dotted and solid black lines show the resulting and expected prices for the LR solution. It is clear that expected prices for the bidding round with minimum cost were higher than final prices, resulting in negative profits for some units in some hours. The difference between the two algorithms is most clear in hour 24, where the resulting price from the BB solution is clearly too low to satisfy the positive profit requirement for Unit 6.

Finally, total revenue for the generators is USD 49259 and USD 49323 for the BB and LR algorithms respectively, showing that the latter method increases generator income, in this case caused by revenues from reserve sales. The difference would have been greater if the LR method would have found the same solution as the BB method, resulting in higher marginal costs and higher prices in hours 3-9 and 21-23.

The fact that the LR algorithm results in higher prices and generator revenues is generally valid, because in this algorithm also startup and operation dependent constant costs are compensated.

5.5.2 Multiple ancillary servic es

A new simulation was performed with the model with basically the same data as in Section 5.5.1, but with the following requirements to primary and tertiary reserves:

primary reserves Ó30 MW
tertiary reserves Ó150 MW

Moreover, the contribution to reserves from each unit were now limited by their ramping rates and by absolute values as given in Appendix A.5. This makes the problem more realistic, but also more difficult to solve.

Table 5-5 shows the resulting unit commitment plan, where differences with Table 5-2, the system with one simple reserve requirement, are underlined. In hours 1-8 and 10-16 it is necessary to have more units on line to satisfy the reserve requirements in this case. Units 2 and 3 contribute to tertiary reserves also when they are not spinning.

Table 5-5: Unit commitment plan test system, multiple ancillary services

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 5-3 shows resulting prices for energy and ancillary services.

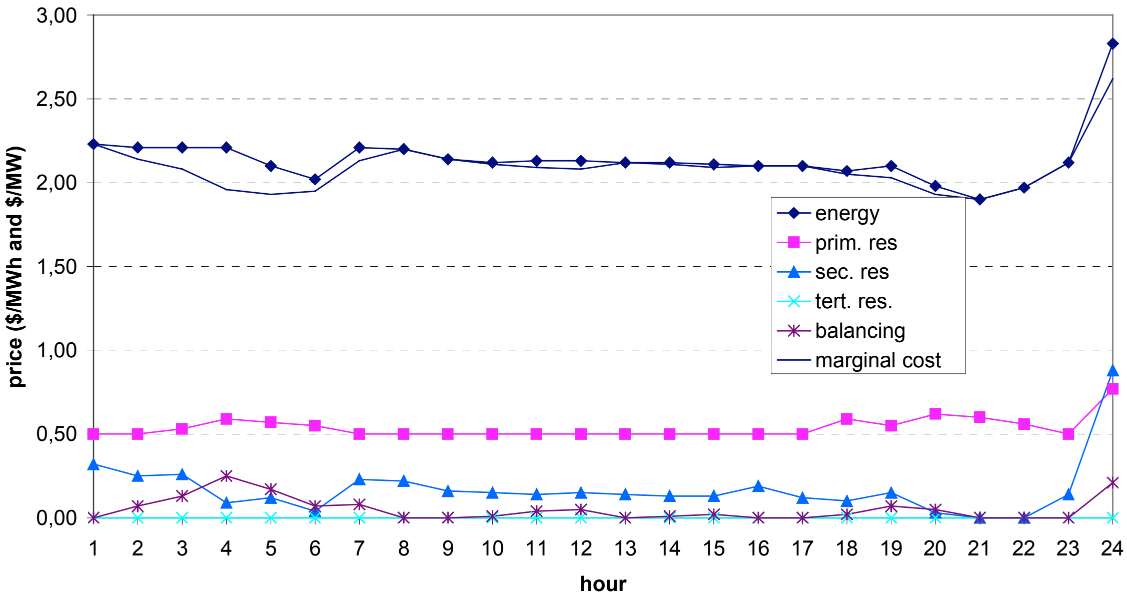

Figure 5-3: Simulated prices, multiple ancillary services

The energy price follows a similar pattern as in Figure 5-2. The price of primary reserve is often equal to its cost, but exceeds the cost in a number of hours because of the opportunity cost effect. The cost of secondary reserves is positive in most hours, while the cost of tertiary reserves is zero for all hours in this case. The price of balancing production is positive in the hours where price is not equal to marginal cost

This example shows that it is possible to solve the UC and dispatch problem for the multiple ancillary services case. Figure 5-3 also shows that price exceeds marginal cost in a number of hours, and that the balancing production mechanism makes it possible to obtain a market-based solution in these hours.

5.5.3 Demand side participatio n

To investigate the effect of demand side participation, 200 MW of demand was defined as inelastic but able to act as secondary or tertiary reserve in the hours 1-9 and 24 at zero marginal cost. In practice, this could be obtained by offering fixed annual payments. Table 5-6 shows the resulting commitment schedule, where differences with Table 5-5 are underlined.

Because of the load participation, the expensive Unit 8 has to run only for one hour, while it is not necessary to start Unit 6 in the last hour. Instead, it is necessary to keep Unit 2 running for all hours. Total simulated cost is USD 47662 and consumer payment USD 50130, respectively USD 300 and USD 1372 lower than without consumer participation. Thus, the result is a net loss of USD 1072 for the generators. The profit for consumers naturally depends on the fixed compensation for being available as reserve and the lost utility of being used as reserve.

Table 5-6: Unit commitment plan test system, demand side participation

hour unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 5-4 shows resulting prices for this case. Energy prices are on average slightly lower, but not much. The greatest difference occurs in hour 24, because it is unnecessary to start Unit

6. As a result, the balancing price is also zero in that hour. Because of the demand side delivery of secondary reserve, its price is zero in the hours 4-9. There is no effect on tertiary reserves, because its price was zero already. However, for cases with higher load, also the price of tertiary reserve is non-zero, and is influenced by demand side participation.

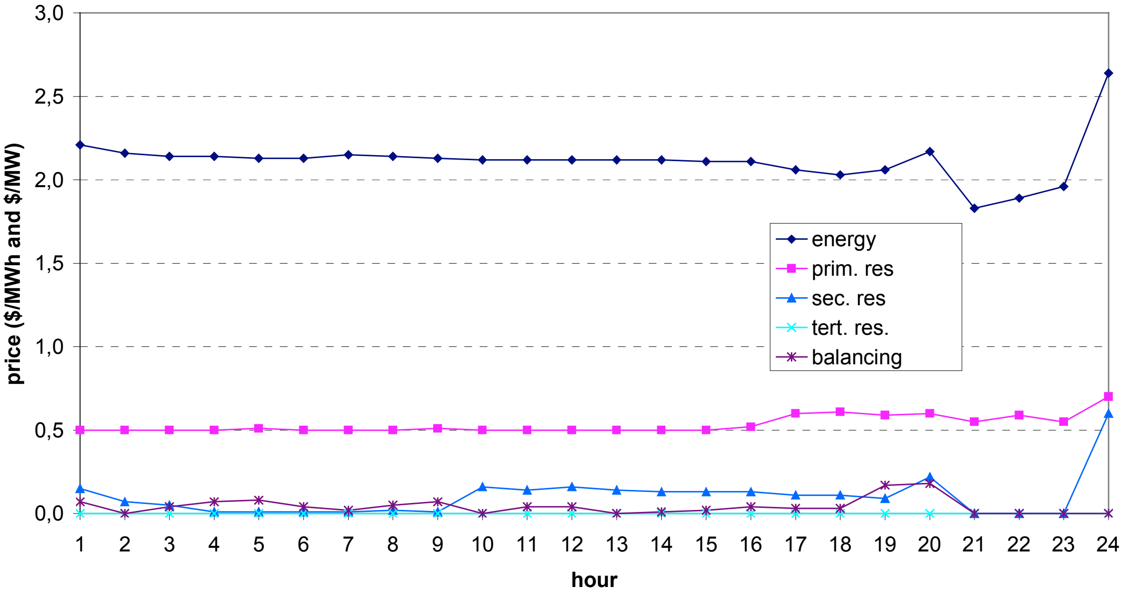

Figure 5-4: Simulated prices with demand side participation

5.6 Ancillary service markets and unit profits

The motivation for analyzing ancillary service markets in the context of this thesis was to find out if such markets increase the profitability of generation capacity. Obviously, generators earn some additional revenues from these services, compared with a situation where they are obliged to deliver them without compensation. However, the commitment decision is subject to a positive profit (because otherwise the maximum of **(5-12)** is zero with u_{it}=0). In cases where this is the binding constraint (which is often the case in the test cases in the previous sections), the energy price has to be higher when there is no payment for ancillary services. So one might suspect that payment for ancillary services in a competitive market lowers the energy price, partly offsetting the increase in generator revenues. To investigate this hypothesis somewhat further, a comparison was made between two cases:

- Case a: includes ancillary service markets like before
- Case b: ancillary services are divided among generators with the same algorithm as before, but generators receive no revenues from them. This can be the situation if for example generators are given some kind of fixed compensation, independent of their actual operation.

These cases were run with 17.5 % higher load than the cases in the previous sections, to emphasize high load conditions. Results are presented in the next table.

Table 5-7: generator profits (USD) with and without compensation for ancillary services

	with compensation	without compensation
energy	60166	60660
primary reserves	449	0
secondary reserves	1019	0
tertiary reserves	159	0
balancing	860	726
sum	62653	61386
cost	56853	56853
generator profits	5800	4523

The results show that the suspected effect occurs. Without compensation for ancillary services, energy prices are higher than with compensation, but no so much as to completely offset the loss of revenue from ancillary services. The detailed hourly results show that energy prices are higher in almost all hours where the positive revenue constraint is binding, but the differences are very small.

Consequently, compensation for ancillary services does increase generator income, but the effect is reduced by somewhat lower energy prices.

5.7 Conclusions on ancillary services markets

In this chapter, a potential solution for an integrated market for energy and ancillary services has been proposed. A model for such a market has been developed, and some test calculations have been done on a 10-unit system.

The unique property of the model is that it recognizes opportunity costs and their influence on prices in a non-convex model. Realistic market prices can be obtained. In principle, the model could be implemented in a real market. However, there are a number of drawbacks:

- The market-clearing algorithm described in Section 5.3 is time consuming. Measures should be taken to improve its performance if it should be used in a real market.
- The type of algorithm used in this model may often give apparently inconsistent results, understandable only for people familiar with the algorithm.
- The model is based on a pool type of market. In such markets, the market operator is concerned with a lot of technical detail, and the whole clearing procedure is opaque for most market participants. This is an impediment for other participants than the traditional generators, and forms an entry barrier.

- In England and Wales, there have been considerable market power problems. The pool structure itself may have been one cause of these problems (although there is no general agreement on this, as discussed other places in this thesis), partly because of the previous point.

The motivation for developing this model in the context of this thesis was to analyse if a market model where generators are explicitly paid for ancillary services increases revenues from investments in capacity. It has been shown in an example that such markets increase generator revenues, but if the market is competitive, this is partly offset by a decrease in the energy price. Together with the remaining uncertainty of the revenues, it is doubtful if the introduction of such markets will increase the willingness to invest in peaking capacity sufficiently to guarantee system security according to traditional measures.

5.8 References

[5.1] F.L. Alvarado, “Methods for the Quantification of Ancillary services in Electric Power Systems”, V SEPOPE Symposium of specialists in electric operational and expansion planning, 19-24 May 1996, Recife, Brazil

[5.2] R.D.Christie, I.Wangensteen, “The Energy Market in Norway and Sweden: Introduction”, Power Engineering Letters in *IEEE Power Engineering Review*, February 1998, Vol. 18, No. 2.

[5.3] “The Markets” [cited 2000-10-30]. Available from Internet <http://www.caiso.com/aboutus/infokit/Markets.html>

[5.4] “Market Rules and Procedures [cited 2000-10-30]. Available from Internet <http://www.iso-ne.com/market_system/MRPnew.html>

[5.5] “Preliminary Report in the Operation of the Ancillary Service Markets of the California Independent System Operator (ISO)”, California Independent System Operator Market Surveillance Committee in Compliance with the July 17, 1998 Order in Docket No. ER98-2843 et al.

[5.6] S. Hao, G.A. Angelidis, H. Singh, A.D. Papalexopoulos, “Consumer Payment Minimization in Power Pool Auctions”, *IEEE Transactions on Power Systems*, Vol. 13, No. 3, August 1998.

[5.7] M. Aganagic, K.H. Abdul-Rahman, J.G. Waight, “Spot Pricing of Capacities for Generation of Reserve in an Extended Poolco Model”, *IEEE Transactions on Power Systems*, Vol. 13, No. 3, August 1998.

[5.8] C.W. Richter, G.B. Sheblé, “Genetic Algoritm Evolution of Utility Bidding Strategies for the Competetive Marketplace”, *IEEE Transactions on Power Systems*, Vol. 13, No. 1, February 1998.

[5.9] E. Hirst, B. Kirby, “Simulating the Operation of Markets for Bulk-Power Ancillary services”, *The Energy Journal*, Vol. 19, No. 3, 1998.

[5.10] A.G. Bakirtzis, "Joint Energy and Rerserve Dispatch in a Competitive Pool using Lagrangian Relaxation", *IEEE Power Engineering Review*, November 1998, Vol. 18, No. 11.

[5.11] Trevor Alvey, Doug Goodwin, Xingwang Ma, Dan Streiffert, David Sun, "A Security-Constrained Bid-Clearing System for the New-Zealand Wholesale Electricity Market", *IEEE Transactions on Power Systems*, Vol. 13, No. 2, May 1998.

[5.12] E.S. Huse, "Power generation scheduling. A free market based procedure with reserve constraints include", PhD Thesis, The Norwegian University of Science and Technology, November 1998.

[5.13] E.S. Huse, I. Wangensteen and H.H. Faanes, "Thermal Power Generation Scheduling by Simulated Competition", *IEEE Transactions on Power Systems*, Vol. 14, No. 2, May 1999.

[5.14] B.H. Bakken, HH. Faanes "Technical and Economic Aspects of operation of using a long submarine HVDC connection for frequency control", *IEEE Transactions on Power Systems*, Vol. 12, No. 3, August 1997.

[5.15] R. Baldick, "The Generalized Unit Commitment Problem", *IEEE Transactions on Power Systems*, Vol. 10, No. 1, February 1995.

[5.16] H.P. Williams, "The economic interpretation of duality for practical mixed integer problems", Survey of mathematical programming: *Proceedings of the 9th international Mathematical Programming Symposium*, Budapest, 1976.

[5.17] Marcelino Madrigal, Victor H. Quintana, Using Optimization Models and Techniques to Implement Electricity Auctions, 2000 IEEE Power Engineering Society Winter Meeting, Singapore , 23-27 January 2000

[5.18] A. Merlin, P. Sandrin, "A New Method for Unit Commitment at Electricité de France", *IEEE Transaction on PAS*, Vol. PA-102, No. 4, May 1983.

[5.19] J.F. Bard, "Short-term Scheduling of Thermal-Electric Generators Using Lagrangian Relaxation", *Operations Research*, Vol. 36, No. 5, September-October 1998.

APPENDIX A.5: TEST SYSTEM DATA

	PMIN	PMAX	A0	$A1_{en}$	A2	$A1_{prim}$	CH	CL	TC	TDOWN	TUP	up/down at t=0
Unit1	15	60	15	1.400	0.0051	0.503	15	123	5	2	3	3
Unit2	20	80	25	1.500	0.0040	0.496	15	123	5	4	3	3
Unit3	30	100	40	1.350	0.0039	0.500	25	110	5	4	4	4
Unit4	25	120	32	1.400	0.0038	0.504	12	100	5	3	4	3
Unit5	50	150	29	1.540	0.0021	0.495	30	130	5	3	1	-3
Unit6	75	280	72	1.350	0.0026	0.510	30	146	6	3	6	-3
Unit7	250	520	105	1.395	0.0013	0.501	60	207	11	4	10	10
Unit8	50	150	100	1.329	0.0014	0.498	80	202	11	2	3	3
Unit9	120	320	49	1.264	0.0029	0.497	50	137	7	5	7	7
Unit10	75	200	82	1.214	0.0015	0.502	70	157	9	6	6	6

	$PMAX_{prim}$	$PMAX_{sec}$	$PMAX_{tert}$	$PMAX0_{sec}$	$PMAX0_{tert}$
Unit1	3	45	45	45	45
Unit2	4	60	60	60	60
Unit3	5	70	70	40	70
Unit4	6	95	95	40	95
Unit5	7.5	75	100	0	0
Unit6	14	30	60	0	0
Unit7	26	30	60	0	0
Unit8	7.5	30	60	0	0
Unit9	16	30	60	0	0
Unit10	10	30	60	0	0

Chapter 6: Capacity subscription

Section 3.1 of this thesis briefly discusses the literature on optimal pricing and investment in regulated power systems. As the literature shows, the central problem of peak load supply is the combination of uncertainty with *ex ante* fixed prices. Many authors conclude that in the optimal solution, occasional rationing will be necessary under extreme conditions.

A central theme in the economic literature on this topic is how to obtain efficient rationing in the optimal-welfare sense. An interesting scheme, called Self-Rationing, was introduced by Panzar and Sibley in 1978 [6.2]. The scheme works as follows: each consumer subscribes to a particular level of capacity *before* "the state of nature is revealed". States of nature are modelled through the stochastic temperature variable t, where a high temperature implies high demand[1]. The consumer pays a capacity charge for the amount he subscribes to, and a usage charge for actual consumption. If his usage exceeds subscribed capacity, a fuse[2] curtails further consumption. Consumers differ according to their willingness to pay for power or, put alternatively, quality of supply, cf. Section 3.4.3. Those with a high willingness to pay for quality of supply would purchase larger fuse sizes than those with a low willingness to pay. In a regulated environment, the task of the utility is to find the optimal usage ($/kWh) and fuse ($/kW) prices, and the optimal installed capacity. Panzar and Sibley derive optimal prices equal to marginal costs of production and capacity respectively, and optimal installed capacity equal to the sum of all fuse sizes. They finally show that their scheme is *ex post* optimal under the assumption that the marginal willingness-to-pay function $P(q,\theta,t)$ where q is the quantity consumed, θ customer type and t temperature (i.e. the stochastic variable), is weakly separable in q and θ, i.e. can be written as $P(h(q,\theta),t)$. The behavioral impact of this assumption is that consumers are similar in the way in which temperature (uncertainty) changes affect their preferences.

In [6.3], Schwarz and Taylor develop a multi-period version of the model, with imperfect correlation of customer demands. They conclude that the energy price should equal marginal cost, but that the fuse price should be less than the marginal capacity cost and installed capacity less than the sum of the fuse sizes. The rationale behind this result is that demand is not perfectly correlated, resulting in maximum demand being less than the sum of the fuse sizes. Furthermore, they state that a single demand charge is generally suboptimal, resulting in difficulties in designing a practical tariff.

[1] Naturally, in a cold climate like the Norwegian, and with electrical heating, a low temperature would imply high demand. In the formal model description a general stochastic variable u will be used instead of t.

[2] This is not a fuse in the usual sense: it will not blow, reducing demand to zero, but limit demand to the fuse size. Here the term fuse will be used for this device.

A remaining problem with both these analyses is that of "untimely curtailments" when idle capacity exists, exemplified in [6.2] by the "insomniac" blowing his fuse at 4 a.m.)[3]. Woo [6.5] recognizes this problem, and proposes a truly innovative solution, *by letting the utility activate the fuses only when a capacity shortage occurs*. Although devised for a regulated environment, this scheme is attractive and feasible in restructured systems for a number reasons:

- Willingness to pay for fuses directly mirrors consumers' preference for quality of supply as discussed in Section 3.4. Consumers with a low preference for quality of supply will obtain small fuses, resulting in considerable curtailment when there is a capacity shortage. On the other hand, consumers with a high preference for quality of supply will obtain large fuses, and will not be curtailed when there is a capacity shortage.
- Capacity installation can be based on consumers' revealed preferences, instead of surveys where consumers answer hypothetical questions.
- Quality of supply becomes a private good instead of a public good, and supply and demand can be matched in an optimal way.
- Introduction of demand for fuses enables the creation of a capacity market, where demand for fuses truly reflects demand for the real function of capacity, i.e. to ensure uninterrupted supply during all "states of nature". If demand is high, capacity prices will become high when there is a shortage, and producers can invest in peaking capacity. On the other hand, is demand is low, i.e. many consumers are satisfied with being curtailed during shortage period, prices will stay low, and no new investments are made.
- Although definitely more complicated than an energy-only market, the transaction costs of self-rationing as proposed by Woo are much lower than e.g. for full spot pricing.
- Once the technological infrastructure for the scheme are in place, network owners might enter into agreements with their customers to employ the same mechanism to alleviate network overloads. Naturally, customers would require compensation, because the probability of being limited by the fuse increases. Agreements like this could be part of advanced network tariffs.

Before supporting these claims, some comments on Woo's paper are appropriate. Based on *ex ante* social welfare optimization for the multi period case, Woo derives optimal prices equal to marginal cost of energy and capacity respectively, and optimal capacity equal to the sum of the fuse sizes. The decisive difference compared with Schwarz and Taylor is the utility's opportunity to control the fuses. Consequently, it will not activate fuses before demand actually equals capacity. Consumers will anticipate this, and obtain fuses according to their expected demand when system demand is at its maximum. Under the same weak

[3] Basically the same criticism is given against demand charges in [6.4] (pp. 69-71), where it is stated that "Demand charges do not send good price signals". At 4 a.m. marginal costs are low, and it economically it makes no sense that consumers reduce their demand to avoid the charge.

separability assumption as Panzar and Sibley, Woo shows that the scheme also is optimal *ex post*.

In spite of its innovative contribution, there are two reasons why inefficiencies remain in Woo's model, as pointed out by Doucet and Roland in [6.6]. Firstly, marginal willingness to pay is generally different between consumers at the time of consumption. This means that once limited by their fuses, some consumers will be willing to pay more for additional consumption than others and they would become better off if mutual trade was possible in some way[4]. Secondly, although the control of fuses reduces this effect, some consumers will still be limited by their consumption while system demand is less than capacity. This can be demonstrated by the simple two-consumer case from [6.6]:

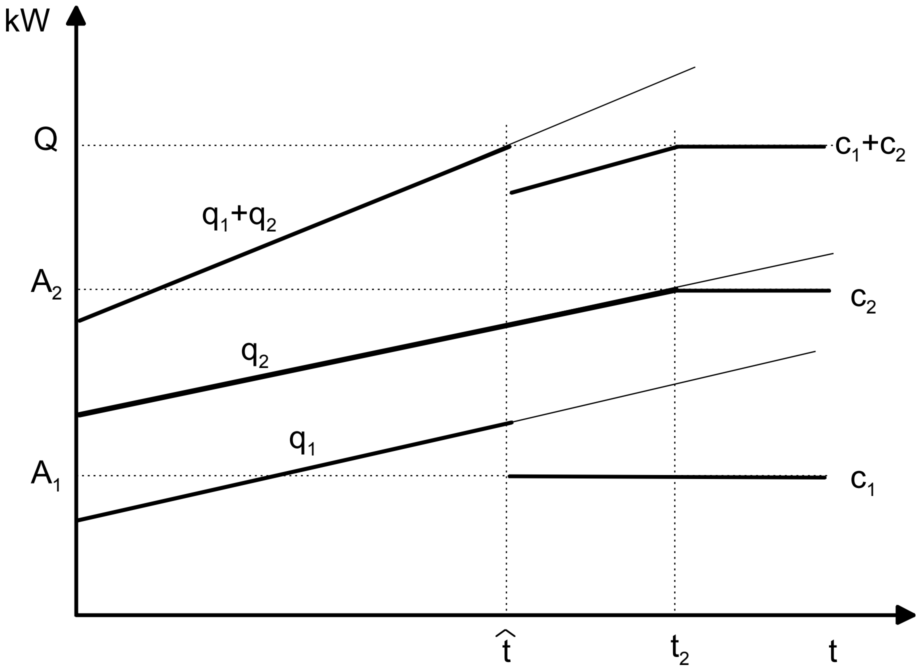

Figure 6-1: An illustration of self-rationing for a two-consumer case

In Figure 6-1, consumer 1 with demand q_1 reaches his fuse size A_1 at a value of t well below $\hat{t}$, the temperature where total demand equals installed capacity Q. Up to this temperature, he is allowed to increase his demand, but at $\hat{t}$ the fuse is activated, and his demand is reduced to his fuse size A_1. Because consumer 2 with demand q_2 has not yet reached his fuse size A_2, total demand drops with an amount equal to $A_2 - q_2(\hat{t})$. In Woo's scheme, this situation is avoided through the property of weak separability, which entails that all consumers reach $\hat{t}$ at the same instant. As Doucet and Roland point out, in that case the possibility to control the fuses is superfluous. However, the assumption of weak separability is unrealistic and consequently too strong.

[4] This point is assessed by Doucet et al. in another paper [6.7]. The scheme proposed there however, depends on a price $\pi(\omega)$ that is contingent on the "state of nature", i.e. it is in reality a spot price.

In [6.6] it is shown that the inefficiencies in Woo's model can be reduced by reducing the use of fuses, i.e. shortening the interval of temperatures over which fuses constrain demand. This can be obtained by increasing the price below $\hat{t}$. A price above marginal cost will result in a new inefficiency. Prices p_0 below and p_1 above $\hat{t}$ and $\hat{t}$ itself (which now becomes a decision variable) should be chosen such as to minimize the total inefficiencies, or more generally to maximize social welfare. Solving the resulting problem, Doucet and Roland find an optimal price p_0 above marginal cost. The optimal price p_1 is equal to marginal cost because above $\hat{t}$, demand is rationed by fuses. The fuse price k is equal to marginal capacity cost minus the shadow price μ of the constraint that the sum of the fuses should be less or equal to installed capacity.

The scheme of Doucet and Roland reduces, but does not eliminate the problem of too much load relief. Furthermore, it will not work in a competitive market, because in this case it will not be possible for a utility or a regulator to fix the price above marginal cost below $\hat{t}$. In the next sections, a model for a competitive market will be derived, and possibilities for practical implementation will be discussed. To minimize efficiency losses, two different strategies to avoid excess load relief are explored:

- random allocation
- optimal allocation according to willingness to pay

Furthermore, the term Capacity Subscription will be used instead of Self-Rationing, because it is more appropriate in a restructured environment: consumers choose to purchase something they want (capacity) not something they do not want (rationing).

6.1 Random allocation of excess capacity

6.1.1 The consumer problem

Initially, the case of random allocation will be described. A formal discussion of the concept of utility in the current context is given in [6.2] or alternatively in [6.8], and will not be repeated here. A "sticky" market price p is assumed within the period N, i.e. it is assumed that the price reflects market conditions over a period of some duration, but can not vary continually like an ideal spot price would[5]. A consumer buys a fuse size A at a per unit price k in order to maximize his expected surplus:

[5] If a spot price would be feasible, there would be no rationing problem.

$$
\begin{aligned}
ES(\theta) = \int_0^N \Bigg\{ & \int_{u_L}^{\hat{u}_n} \int_0^{q_n(p,u_n,\theta)} \big(P_n(q',u_n,\theta) - p\big)\, dq' dF + \\
& \int_{\hat{u}_n}^{u_H} \Bigg[\int_0^{c_n(A(\cdot),p,u_n,\theta)} \big(P_n(q',u_n,\theta) - p\big)\, dq' + \\
& \int_{c_n(A(\cdot),p,u_n,\theta)}^{q_n(p,u_n,\theta)} \big(P_n(q',u_n,\theta) - p\big) \cdot Pr_e(p,k,u_n)\, dq' \Bigg] dF \Bigg\} dn - \\
& k \cdot A(p,k,PR_e,\theta)
\end{aligned}
\qquad \textbf{(6-1)}
$$

Here $P_n(q,u_n,\theta)$ is consumer θ 's marginal willingness to pay for a quantum q at time n and state of nature u_n. P_n is assumed decreasing in q and increasing in u and θ. The stochastic variable u may have values between u_L and u_H, and has a density function *f(u)*, with a cumulative *F(u)*[6]. $\hat{u}_n$ is the value of u_n where total demand is equal to installed capacity:

$$\int_{\theta_L}^{\theta_H} q_n(p,\hat{u}_n,\theta) dG = Q$$

where G is the cumulative consumer distribution function. $c_n(A(\cdot),p,u_n,\theta)$ is given by:

$$c_n(A(\cdot),p,u_n,\theta) = \begin{cases} A(p,k,PR_e,\theta) \text{ if } q_n(p,u_n,\theta) > A(\cdot) \text{ and } u_n \geq \hat{u}_n \\ q_n(p,u_n,\theta) \quad \text{otherwise} \end{cases}$$

The function A is written as $A(p,k,Pr_e,\theta)$ since Pr_e is taken as a parameter by the consumer and $A(\cdot)$ is defined as the optimal choice of fuse by the consumer. It is natural that the consumer receives information about the probability of (non) curtailment from e.g. the system operator.

In **(6-1)**, the first line represents the consumer surplus for values of $u_n < \hat{u}_n$, i.e. before fuses are activated. The second line represents the surplus for values of $u_n \geq \hat{u}_n$, after activation of the fuses. The third line represents the benefit of randomly allocated demand.

With random allocation of excessively fuse-limited demand, the probability for consumer θ of being able to satisfy his excess demand $q_n(p,u_n,\theta) - A(p,k,\theta)$ is equal to unused fuse capacity divided by total desired demand over fuse size:

[6] Woo and Doucet et al. use the temperature variable *t* to describe uncertainty, and assume that demand increases with temperature. Here a more general approach is taken.

$$\Pr_e(p,k,u_n)=\frac{\int_{\theta_L}^{\theta_H}(1-I(A(\cdot),p,u_n,\theta))(A(p,k,\theta)-q_n(p,u_n,\theta))dG}{\int_{\theta_L}^{\theta_H}I(A(\cdot),p,u_n,\theta)(q_n(p,u_n,\theta)-A(p,k,\theta))dG}\qquad\textbf{(6-2)}$$

where

$$I(A(\cdot),p,u_n,\theta)=\begin{cases}1 & \text{if } q_n(p,u_n,\theta)>A(\cdot) \text{ and } u_n>\hat{u}_n\\ 0 & \text{otherwise}\end{cases}$$

The numerator designates subscribed capacity not used at u_n, while the denominator equals desired demand minus fuses sizes for those customers where demand is limited by their fuses. Provided all excess load relief is allocated, the effective system demand at $u_n \geq \hat{u}_n$ is:

$$\int_{\theta_L}^{\theta_H} A(\cdot)dG=S\qquad\textbf{(6-3)}$$

The optimal subscribed capacity chosen by the consumer satisfies the first order condition:

$$\frac{\partial ES(\theta)}{\partial A}=$$
$$\int_{0}^{N}\int_{\hat{u}_n}^{u_H} I(A(\cdot),p,u_n,\theta)(P_n(A(\cdot),u_n,\theta)-p)(1-Pr_e(\cdot))dFdn-k=0\qquad\textbf{(6-4)}$$

This equation states that the expected marginal consumer benefit in the optimum should equal the marginal cost of acquiring an additional unit of fuse capacity. The expected marginal benefit is reduced by the possibility of being randomly allotted excessively relieved load when desired demand exceeds fuse size.

Using the Envelope Theorem, the derivatives of **(6-1)** with respect to p and k are given in Appendix A6.1.

6.1.2 The producer problem

A constant marginal cost of generation b and capacity cost β are assumed. Furthermore, full allocation of empty capacity at $u_n \geq \hat{u}_n$ implies that demand equals available capacity in these states. Thus producer surplus is, like in [6.5]:

$$E\pi = (p-b)\int_0^N \left\{ \int_{\theta_L}^{\theta_H} \int_{u_L}^{\hat{u}_n} q_n(p,u_n,\theta)dFdG + \int_{\theta_L}^{\theta_H} [1-F(\hat{u}_n)]A(p,k,\theta)dG \right\} dn + kS - \beta Q \quad \textbf{(6-5)}$$

The following constraints apply:

$$Q \geq S \qquad 0 \leq n \leq N,\ \hat{u}_n < u \leq u_H \quad \textbf{(6-6)}$$

6.1.3 Welfare-optimal solution with random allocation of excess capacity

The purpose of this section is to find optimal prices and installed capacity from a single-owner point of view. The question will then be if this solution is feasible in a deregulated market.

Based on the previous two sections, we can set up the Lagrangian for the problem:

$$\mathscr{L} = \int_{\theta_L}^{\theta_H} ES(\theta)dG + E\pi + \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un}(Q-S)dFdn \quad \textbf{(6-7)}$$

Here the α_{un} are the Lagrange multipliers of the constraints **(6-6)**. The derivatives of **(6-7)** are given in Appendix A6.1.

Solutions with *p* or *Q* equal to zero are of no practical interest, if they should exist. A solution with *k* equal to zero does not make sense either, for then consumers would procure fuses that were large enough for any conceivable demand, and there would be no self-rationing. Thus an inner solutions is sought, satisfying the KKT-conditions:

$$\frac{\partial \mathscr{L}}{\partial p} = 0,\ \frac{\partial \mathscr{L}}{\partial k} = 0,\ \frac{\partial \mathscr{L}}{\partial Q} = 0.$$

The resulting prices can in principle be calculated, but become somewhat cumbersome to express. Generally, in the optimal state, prices are not equal to marginal cost, and the *ex ante* optimal solution will not be obtained in a market environment. Neither will the solution be optimal *ex post*, because the random allocation mechanism naturally leaves consumers with different willingness to pay, cf. [6.5], page 76.

6.2 Optimal allocation of ex cess capacity

An alternative to random allocation of excessive load relief is optimal allocation through prices, i.e. the idle capacity resulting from fuse activation at $u_n = \hat{u}_n$ is allocated to the consumers with highest willingness to pay. In practice, it is necessary for consumers to be continually informed about prices (when $u_n > \hat{u}_n$), and metering equipment must be in place for continuous or at least hourly metering. Once this equipment is in place, one might ask why spot pricing should not be used. There are several good reasons for this:

- Continuous spot pricing (8760, 17520 or even more periods per year) for all consumers necessitates the handling of enormous amounts of data. This is because although the spot price always equals *b* in the model, it will vary continuously in the real world. Spot pricing only when there is a capacity shortage reduces this amount to just a few percents. Because of this, cheaper solutions can be applied, both locally at the consumer's place, centrally and with respect to communication.
- Consumers may not be interested in spot pricing, but may value the option of being able to buy electricity in excess of their *ex ante* obtained fuse size when a real shortage situation occurs.
- This form for "contingent spot pricing" may also solve the transaction cost problem of spot pricing. The cost of installing the additional equipment to buy power on spot conditions during capacity shortage could be borne by those consumers who value this option most.

Under these assumptions, *ES(θ)* and *Eπ* become:

$$\begin{aligned} ES(\theta) = \int_0^N \Bigg\{ & \int_{u_L}^{\hat{u}_n} \int_0^{q_n} \left(P_n(q',u_n,\theta) - p_0\right) dq' dF + \\ & \int_{\hat{u}_n}^{u_H} \Bigg\{ \int_0^{c(A(\cdot),p,u_n,\theta)} \left(P_n(q',u_n,\theta) - p_0\right) dq' + \\ & \int_{c(A(\cdot),p,u_n,\theta)}^{q_n(p_n(u_n),u_n,\theta)} \left(P_n(q',u_n,\theta) - p_n(u_n)\right) dq' \Bigg\} dF \Bigg\} dn - \\ & kA(p,k,PR_e,\theta) \end{aligned} \qquad \textbf{(6-8)}$$

where p_0 is the optimal price when $u_n < \hat{u}_n$ and $p_n(u_n)$ the spot price otherwise.

$$\begin{aligned} E\pi = \int_0^N \int_{\theta_L}^{\theta_H} \Bigg\{ (p_0 - b) \Bigg[& \int_{u_L}^{\hat{u}_n} q_n(p_0,u_n,\theta) dF + \\ & \int_{\hat{u}_n}^{u_H} (1 - I(\cdot)) q_n(p_0,u_n,\theta) dF \Bigg] + \\ & \int_{\hat{u}_n}^{u_H} (p_n(u_n) - b) I(\cdot) q_n(p_n(u_n),u_n,\theta) dF \Bigg\} dG dn + kS - \beta Q \end{aligned} \qquad \textbf{(6-9)}$$

The Lagrangian formally has the same definition as in **(6-7)**, and the derivatives are again given in Appendix A.6.1.

The set of equations describing the KKT conditions can by inspection be shown to satisfy:

$$p_0 = b, \quad k = \beta = \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un} dF dn, \quad Q = S \tag{6-10}$$

and

$$p_n(u_n) = b - \frac{(I(\cdot)-1)q_n(p_n(u_n),t,\theta) + A(p_0,k,\theta)}{\partial q_n / \partial p_n} \tag{6-11}$$

In this case p_0 and k are equal to marginal costs. $p_n(u_n)$ equals marginal cost *plus* (because $\partial q_n / \partial p_n < 0$) the amount necessary to limit demand to capacity.

Because optimal prices equal marginal costs, capacity subscription with optimal allocation of excess load relief can in principle result in an optimal solution in a deregulated market structure.

6.3 Implementation issues

The derivations so far do not regard necessary capacity for reserves and losses. The impact of reserve requirements can be viewed in different ways. One viewpoint is that once suppliers have sold a certain amount of capacity, it is their responsibility to have that amount available during peak load conditions. Because the availability of their generators is less than 100 %, they cannot sell capacity equalling their installed capacity. They have to sell a lower amount and in addition should obtain the right to buy capacity from other in case of outages and/or insure the risk of having to pay penalty payments due to failure to supply. Another viewpoint is that a power system can not operate without some level of available reserves. Because of the reserve requirement, it is not sufficient that available capacity equals demand. Instead, available capacity should equal demand plus the reserve requirement, i.e. instead of **(6-6)**:

$$Q \geq S \cdot (1+r) \qquad 0 \leq n \leq N,\ \hat{u}_n < u \leq u_H \tag{6-12}$$

where r is the relative reserve requirement. This also influences the definition of $\hat{u}_n$, which becomes the value of u_n where total demand *plus reserve requirements* is equal to installed capacity:

$$(1+r) \cdot \int_{\theta_L}^{\theta_H} q_n(p,\hat{u}_n,\theta) dG = Q$$

The result of this is that (1+r)·S substitutes S in the equations. Essentially optimal capacity Q in **(6-10)** equals (1+r)·S. Moreover the optimal price of capacity k in **(6-10)** also increases with the factor 1+r. With a constant price b for capacity this result is intuitively correct: if a reserve requirement is imposed on the suppliers, their costs will increase with the same factor as the reserve requirement. If the market is competitive, this cost increase should be reflected in the market price.

The latter viewpoint, to impose a reserve requirement on the sellers of capacity, is probably most suitable for primary reserves because of the short time frame. This is related to the question if capacity is to be perceived as being a (half)hourly average or a momentary value. The former viewpoint can be applicable to reserves with a longer time frame.

Losses are a special form of "demand" that must be treated separately in this context. It would be natural to think of losses as a demand by the System Operator (to be recovered by transmission tariffs), for which the System Operator could buy the necessary capacity. However, it is of course not possible to limit losses to some maximum level, for which the System Operator should have bought "fuses". One solution would be to treat losses in a similar way as reserves, as shown above. In this case, a requirement would be imposed on suppliers to supply subscripted capacity plus an estimated loss percentage. The System Operator would have to make e.g. annual estimates of losses during peak demand.

Another issue is the distribution of capacity across the transmission network. Ideally, self-rationing would guarantee sufficient capacity to be available, provided generators have not "oversold". However, if capacity is locked in behind a transmission bottleneck, it will not be available to the consumer. This could for example be solved with a price area mechanism, similar to the system used in the present Norwegian spot market.

These issues are not pursued further in the present context, but must be given due consideration in the case of practical implementation

6.4 The use of capacity subs cription in restructured power systems

Capacity subscription in a restructured power system enables the creation of a true capacity market, where "capacity during system-peak conditions" is bought and sold. The essential feature of the solution is a centrally controlled "fuse" device, enabling the system operator to limit demand to subscribed capacity when generation capacity is insufficient to serve unconstrained demand. Based on their preferences for uninterrupted supply and the price of capacity, consumers buy their preferred amount of capacity, while generators can sell their available capacity. The interesting feature of this solution is, that it guarantees that there always will be "enough" generating capacity, under the assumption that generators do not oversell, against which necessary measures can be taken.

Because not all consumers will reach their subscribed capacity at the instant when the limits are activated, there will be some idle capacity after activation, representing an economic inefficiency. Two methods have been analysed to avoid this inefficiency: random allocation and optimal (price based) allocation of idle capacity. With random allocation, optimal prices are generally not equal to marginal prices, with the result that a competitive market will not find this optimum. With price based allocation, optimal prices are equal to marginal costs for energy and capacity, and the solution is feasible in a competitive market.

Price based allocation of idle capacity requires (half)hourly metering. It can be left to the market participants to obtain metering devices: consumers that put a high value on the possibility of being able to buy electricity in excess of their subscribed capacity during system-peaks, will invest in such devices, while others won't. Typically, this will be more interesting for large consumers than for smaller ones. In addition to this feature, the capacity subscription scheme requires only a moderately expensive infrastructure: the “fuse” device and the communication opportunity that allows the system operator to control the device. It can be left to the market to develop solutions to control demand efficiently at the consumer’s place. If capacity prices are high, there will be a motivation to develop devices that control demand with minimal loss of comfort, allowing consumers to buy smaller “fuses”. On the other hand, if there is excess capacity, prices will be low, and consumers will prefer to subscribe on higher levels of capacity.

Mainly from a theoretical point of view, this chapter has shown that the capacity subscription scheme can solve the peak-capacity problem in restructured power systems. The promising results invite to further research in the direction of practical implementation.

6.5 References

[6.1] Michael A. Crew, Chitrus S. Fernando, Paul R. Kleindorfer, “The Theory of Peak-Load Pricing: A Survey”, *Journal of Regulatory Economics*, Vol. 8, No. 3, 1995, pp. 215-248.

[6.2] J.C. Panzar, D.S. Sibley, “Public Utility Pricing under Risk: the case of self-rationing”, *The American Economic Review*, Vol. 68, No. 5, 1978, pp. 888-895.

[6.3] Peter M. Schwarz, Thomas N. Taylor, “Public Utility Pricing under Risk: the Case of Self-Rationing: Comments and Extension”, *The American Economic Review*, Vol. 77, No. 4, 1987, pp. 734-739.

[6.4] Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, Roger E. Bohn, “Spot Pricing of Electricity”, Kluwer Academic Publishers, 1988.

[6.5] Chi-Keung Woo, “Efficient Electricity Pricing with Self-Rationing”, *Journal of Regulatory Economics*, Vol. 2, 1990, pp. 69-81.

[6.6] Joseph A. Doucet, Michel Roland, “Efficient Self-Rationing of Electricity Revisited”, *Journal of Regulatory Economics*, Vol. 5, 1993, pp. 91-100.

[6.7] Joseph A. Doucet, Kyung Jo Min, Michel Roland, Todd Straus, “Electricity rationing through a two-stage mechanism”, *Energy Economics*, Vol. 18, pp. 247-263, 1996.

[6.8] Seong-Uh Lee, “Welfare-optimal Pricing and Capacity Selection Under An Ex Ante Maximum Demand Charge”, *Journal of Regulatory Economics*, Vol. 5, 1993, pp. 317-335.

APPENDIX A.6: DERIVATIONS AND EXAMPLE

A6.1 Derivations

Derivatives of *Pr_e*

The derivative of Pr_e **(6-2)** with respect to p is:

$$\frac{\partial Pr_e}{\partial p} = \frac{\Omega_1}{\Omega_2} \quad \text{where}$$

$$\begin{aligned}\Omega_1 = &\int_\theta (1-I)\left(\frac{\partial A}{\partial p} - \frac{\partial q}{\partial p}\right) dG \int_\theta I(q_n - A)\,dG - \\ &\int_\theta (1-I)(A - q_n)\,dG \int_\theta I\left(\frac{\partial q}{\partial p} - \frac{\partial A}{\partial p}\right) dG \quad \text{and}\end{aligned} \qquad \textbf{(A6-1)}$$

$$\Omega_2 = \left(\int_\theta I(q - A)\,dG \right)^2$$

where the arguments of I, A and q have been omitted for readability. I and $(1\text{-}I)$ are positive, $\frac{\partial A}{\partial p}$ can be shown to negative (cf. [6.8]), $\frac{\partial q}{\partial p} < 0$ by assumption. This leaves an ambiguity with respect to the sign of the first integral. However, it is plausible to assume that the fuse size's sensitivity on p, which is a kind of second-order effect, is less than the actual consumption's sensitivity on p, and with this assumption the first integral is positive. $q_n \geq A$ when I=1, and consequently the first term of $\Omega_1 > 0$. When $(1\text{-}I) > 0$ and $u_n > \hat{u}_n$ (because the derivative is taken within an integral from $\hat{u}_n$ to u_H, cf.**(6-1)**), $A \geq q_n$, and consequently the second term of $\Omega_1 < 0$ and thus $\Omega_1 > 0$. Because also $\Omega_2 > 0$, $\frac{\partial Pr_e}{\partial p} > 0$. The intuitive explanation of this is that a higher price p reduces demand, thus increasing the probability of randomly receiving power when $u_n > \hat{u}_n$.

The derivative of Pr_e **(6-2)** with respect to k is:

$$\frac{\partial Pr_e}{\partial k} = \frac{\int_\theta (1-I)\frac{\partial A}{\partial k}\,dG \int_\theta I(q_n - A)\,dG + \int_\theta (1-I)(A - q_n)\,dG \int_\theta I\frac{\partial A}{\partial k}\,dG}{\left(\int_\theta I(q - A)\,dG \right)^2} \qquad \textbf{(A6-2)}$$

This expression is < 0, because $\frac{\partial A}{\partial k} < 0$, $q_n \geq A$ when I=1 and A ≥ q_n when I=0.

Derivatives of expected consumer surplus, random allocation

$$\begin{aligned}\frac{\partial ES(\theta)}{\partial p} = &-\int_0^N \int_{u_L}^{u_H} c_n(A(\cdot),p,u_n,\theta)dFdn - \\ &\int_0^N \int_{\hat{u}_n}^{u_H} (q_n(p,u_n,\theta)-c_n(A(\cdot),p,u_n,\theta))\cdot Pr_e(p,k,u_n)dFdn + \\ &\int_0^N \int_{\hat{u}_n}^{u_H} \int_{c_n(A(\cdot),p,u_n,\theta)}^{q_n(p,u_n,\theta)} \left\{(P_n(q',u_n,\theta)-p)\frac{\partial Pr_e}{\partial p}\right\}dq'dFdn + \\ &\int_0^N (1-PR_e(p,k,\hat{u}_n))\frac{\partial \hat{u}_n}{\partial p}\int_{c_n(A(\cdot),p,\hat{u}_n,\theta)}^{q_n(p,\hat{u}_n,\theta)}(P_n(q',\hat{u}_n,\theta)-p)dq'dn\end{aligned} \qquad \textbf{(A6-3)}$$

The first term represents the effect on surplus due to "planned" (i.e. non-randomly allocated) consumption. The second term gives the direct effect of a price change on randomly allocated consumption, while the third gives the effect of the price change through the probability of being allocated excessively limited capacity. The last term represents the effect of a price change on surplus through its effect on the instant $\hat{u}_n$ of the stochastic variable u_n where fuses are activated.

$$\begin{aligned}\frac{\partial ES(\theta)}{\partial k} = &-A(p,k,PR_e,\theta) + \\ &\int_0^N \int_{\hat{u}_n}^{u_H} \int_{c_n(A(\cdot),p,u_n,\theta)}^{q_n(p,u_n,\theta)} \left\{(P_n(q',u_n,\theta)-p)\frac{\partial \mathrm{Pr}_e}{\partial k}\right\}dq'dFdn - \\ &\int_0^N PR_e(p,k,u_n)\frac{\partial \hat{u}_n}{\partial k}\int_{c_n(A(\cdot),p,\hat{u}_n,\theta)}^{q_n(p,\hat{u}_n,\theta)}(P_n(q',\hat{u}_n,\theta)-p)dq'dn\end{aligned} \qquad \textbf{(A6-4)}$$

The first term again represents the direct effect of a price change on planned consumption, and second the effect of a change in k on the probability of being allocated excessively limited capacity. The last term represents the effect of a change in k on surplus through its effect on the instant $\hat{u}_n$ of the stochastic variable u_n where fuses are activated.

Derivation of $\mathscr{L}$, random allocation

With random allocation (*ES(θ)* given by **(6-1)** and *Eπ* by **(6-5)**), the derivatives of **(6-7)** with respect to *p*, *k* and *Q* are given by:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial p} &= (p-b)\cdot \int_0^N \int_{\theta_L}^{\theta_H} \left\{ \int_{u_L}^{\hat{u}_n} \frac{\partial q_n}{\partial p} dF + \int_{\hat{u}_n}^{u_H} \frac{\partial A}{\partial p} dF + \right. \\
&\quad \left. \big(q_n(p,\hat{u}_n,\theta) - A(p,k,\theta))\big) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial p} \right\} dGdn + k \cdot \frac{\partial S}{\partial p} - \\
&\quad \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un} \frac{\partial S}{\partial p} dFdn - \int_0^N \alpha_{\hat{u}_n n} \cdot (Q-S) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial p} dn + \\
&\quad \int_0^N \int_{\theta_L}^{\theta_H} \int_{\hat{u}_n}^{u_H} \int_{c_n(A(\cdot),p,u_n,\theta)}^{q_n(p,u_n,\theta)} \left\{ (P_n(q',u_n,\theta) - p) \frac{\partial Pr_e}{\partial p} \right\} dq' dFdGdn + \\
&\quad \int_0^N \int_{\theta_L}^{\theta_H} (1 - PR_e(p,k,\hat{u}_n)) \frac{\partial \hat{u}_n}{\partial p} \int_{c_n(A(\cdot),p,\hat{u}_n,\theta)}^{q_n(p,\hat{u}_n,\theta)} (P_n(q',\hat{u}_n,\theta) - p) dq' dGd
\end{aligned} \tag{A6-5}$$

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial k} &= (p-b)\cdot \int_0^N \int_{\theta_L}^{\theta_H} \left\{ \int_{\hat{u}_n}^{u_H} \frac{\partial A}{\partial k} dF + \right. \\
&\quad \left. (q_n(p,\hat{u}_n,\theta) - A(p,k,\theta)) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial k} \right\} dGdn + \\
&\quad k \cdot \frac{\partial S}{\partial k} - \int_0^N \left\{ \int_{\hat{u}_n}^{u_H} \alpha_{un} \cdot \frac{\partial S}{\partial k} dF - \alpha_{\hat{u}_n n} \cdot (Q-S) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial k} \right\} dn + \\
&\quad \int_0^N \int_{\theta_L}^{\theta_H} \int_{\hat{u}_n}^{u_H} \int_{c_n(A(\cdot),p,u_n,\theta)}^{q_n(p,u_n,\theta)} \left\{ (P_n(q',u_n,\theta) - p) \frac{\partial \mathrm{Pr}_e}{\partial k} \right\} dq' dFdn - \\
&\quad \int_0^N \int_{\theta_L}^{\theta_H} PR_e(p,k,u_n) \frac{\partial \hat{u}_n}{\partial k} \int_{c_n(A(\cdot),p,\hat{u}_n,\theta)}^{q_n(p,\hat{u}_n,\theta)} (P_n(q',\hat{u}_n,\theta) - p) dq'dn
\end{aligned} \tag{A6-6}$$

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial Q} &= (p-b) \cdot \frac{\partial S}{\partial Q} \cdot (1 - F(\hat{u}_n)) + k \cdot \frac{\partial S}{\partial Q} - \\
&\quad \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un} \cdot \frac{\partial S}{\partial Q} dFdn - \beta + \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un} dFdn - \int_0^N \frac{\partial \hat{u}_n}{\partial Q} dn
\end{aligned} \tag{A6-7}$$

Derivation of $\mathscr{L}$, optimal allocation

With optimal allocation (*ES(θ)* given by **(6-8)** and *Eπ* by **(6-9)**), the derivatives of **(6-7)** become:

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial p_0} = (p_0 - b)\int_0^N \int_{\theta_L}^{\theta_H} \Bigg\{ & \int_{u_L}^{\hat{u}_n} \frac{\partial q_n}{\partial p_0} dF + \int_{\hat{u}_n}^{u_H} \frac{\partial A}{\partial p_0} dF + \\ & \big(q_n(p_0, \hat{u}_n, \theta) - A(p_0, k, \theta)\big) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial p} \Bigg\} dGdn + \\ & k\frac{\partial S}{\partial p_0} - \int_0^N \int_{\hat{u}_n}^{u_H} \alpha_{un} \frac{\partial S}{\partial p_0} dFdn - \int_o^N \alpha_{\hat{u}_n n}(Q - S)\frac{\partial \hat{u}_n}{\partial p_0} dn\end{aligned} \qquad \textbf{(A6-8)}$$

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial k} = -\int_0^N \int_{\theta_L}^{\theta_H} & \big[(p_0 - b)(1 - I(\cdot))q_n(p_0, \hat{u}_n, \theta) + \\ & (p_n(\hat{u}_n) - b)I(\cdot)q_n(p_n(\hat{u}_n), \hat{u}_n, \theta)\big] \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial k} dGdn + k\frac{\partial S}{\partial k} - \\ & \int_0^N \left\{ \int_{\hat{u}_n}^{u_H} \alpha_{un} \frac{\partial S}{\partial k} dF - \alpha_{\hat{u}_n n}(Q - S) \cdot f(\hat{u}_n) \cdot \frac{\partial \hat{u}_n}{\partial k} \right\} dn\end{aligned} \qquad \textbf{(A6-9)}$$

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial p_n} = \int_0^N \int_{\theta_L}^{\theta_H} \int_{\hat{u}_n}^{u_H} & \big[(I(\cdot) - 1)q_n(p_n(u_n), u, \theta) + A(p_0, k, \theta) + \\ & (p_n(u_n) - b)\frac{\partial q_n}{\partial p_n}\bigg] dFdGdn\end{aligned} \qquad \textbf{(A6-10)}$$

and $\frac{\partial \mathscr{L}}{\partial Q}$ like in **(A6-7)**.

A6.2 Capacity Subscription and Market Operation - A Simple Example

The interaction between market operation and capacity subscription will be illustrated by a simple example. The market organisation is similar to the Nordic market, with a 24-hour ahead (spot) market, and an additional regulating power market. In the example, there are two generators, each having bilateral contracts. There are no other participants in the market. Reserve requirements are neglected, and generator availability is assumed to be 100 %. The relevant characteristics of the generators are given in the following table.

generator / demand	installed capacity (MW)	bilateral demand (MW)	subscribed capacity (MW)
G1 / D1	100	90	100
G2 / D2	100	120	100

1. Capacity shortage is not foreseen before spot market clearing

At this stage, it is up to the Market Operator (probably assisted by the System Operator) to forecast the necessity of fuse activation. This situation may occur because the forecasts made by the generators on the one hand, and the Market or System Operator on the other hand may not be identical.

For the particular hour, G1 sells 0 MW at a price of 20 and 10 MW at a price of 30.

a) G2 buys 20 MW at a price of 30 or lower and 0 MW at a higher price.

This means G2 gambles on either fuse activation or being able to buy the deficiency cheaper on the regulating power market. The figure below shows the market clears with a price of 30. G1 sells 10 MW and G2 buys 10 MW.

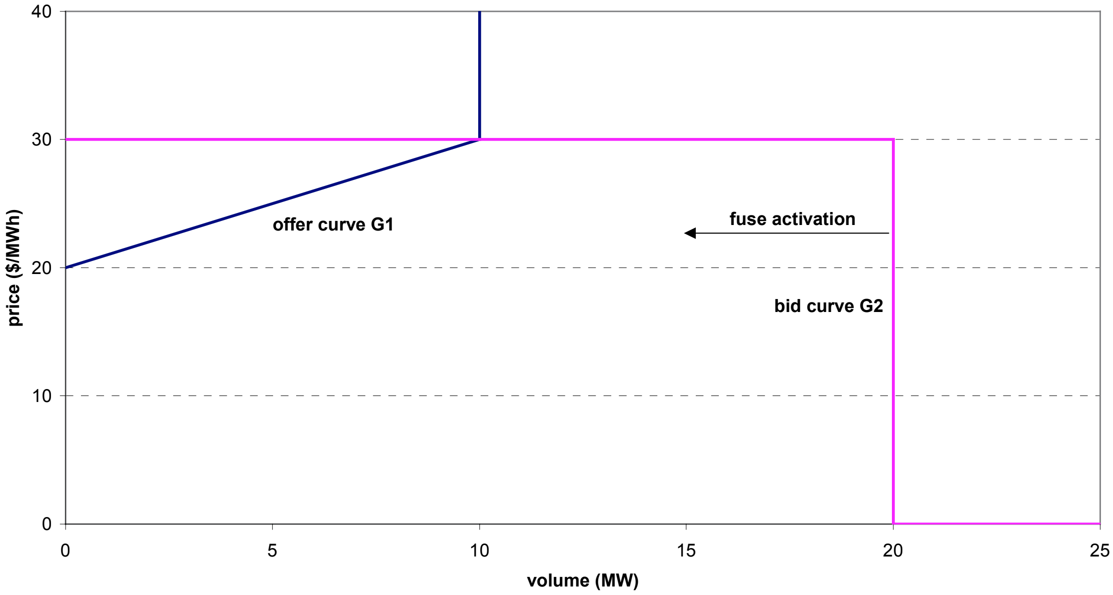

Figure A6-2: Market cross determination, case 1.a

At some instance after spot market clearing, the System Operator decides fuse activation is necessary. Initially, demand is limited to 90+100 = 190 MW. If the random allocation procedure is used, 10 of the 20 MW rationed D2 will be allocated again. D2 becomes effectively 110 MW, and there is no need for regulation. If price based allocation is used, D2 is again rationed to 100 MW. Because of this, G2 has a surplus of 10 MW to sell on the regulating power market, which can be bought by D2. In the figure, it is indicated how the activation of the fuses in reality moves the bid curve of G2 inward.

b) G2 buys 20 MW at any price

There is no market cross, as shown in the next figure. The Market Operator will conclude that it is necessary to activate fuses, and a new round of spot market clearing is carried out.

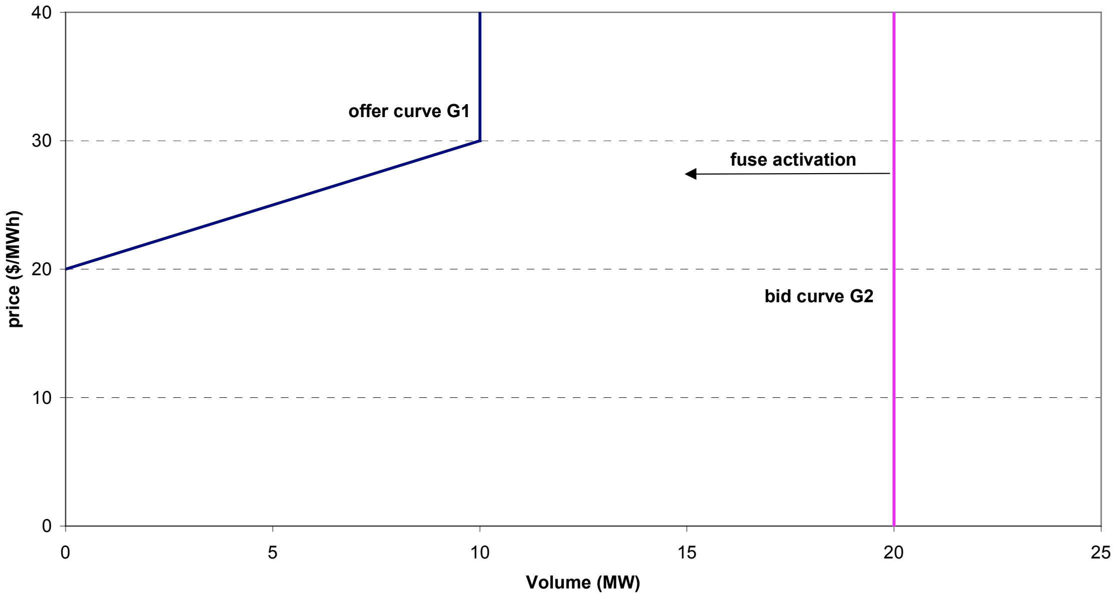

Figure A6-3: Market cross determination, case 1.b

2. Capacity shortage is foreseen before spot market clearing

This can occur either after 1.b or because the demand forecasts by the System or Market operator indicate this situation anyway. G1 estimates his local demand to 90 MW, like before. With random allocation, G2 cannot be certain of his final demand. In this case the bidding behaviour may be like in 1.a. With price based allocation, G2 knows that D2 will not exceed 100 MW. It is now up to D2 to place bids on buying back his rationed demand. Depending on the market organization, this may be done in the spot or regulating power market.

Chapter 7: AVOIDING A CAPACITY DEFICIENCY – A POLICY ANALYSIS

The discussions in Chapter 2 have shown that if left to itself, a market where generators' revenues are based on marginal cost only, is probable not to provide the generation capacity to supply peak-demand of short duration with acceptable security. As a result, a capacity deficiency resulting in a certain amount of random rationing can be expected. This is expensive and socially unacceptable. The problem has been acknowledged in several market designs, and a number of policies have been implemented to handle this problem. In this chapter, six different policies or models will be analysed. Five of these are actually in use in various systems, while the sixth is the capacity subscription model from the previous chapter.

In this chapter, first a number of criteria for comparison between policies will be described. Next, the five implemented policies will be described and briefly discussed. In Section 7.3, the various policies will be discussed and compared based on the criteria, and a conclusion will be drawn.

7.1 Criteria for policy analysis

When comparing alternative strategies to attain societal objectives, Harvey Averch [7.1] uses ten criteria or factors. The main strategies compared are "markets" and "bureaus", a generic description of regulated monopolies. These criteria are highly relevant for comparing various solutions for the peak-load capacity problem. In the following, the criteria will be briefly described, closely following [7.1], at the same time defining their relevance with respect to the current problem.

Static efficiency

Static efficiency refers to the welfare-optimal match of consumption and supply, according to classical economics obtained when marginal cost equals marginal willingness to pay. This is the typical welfare-analysis criterion, that is given much weight in economic analysis. But as stated in [7.1], "… decision makers have little patience with considerations of efficiency, and these are never given as much weight in actual decisions as economists believe they should have".

Dynamic efficiency

Firms obtain static efficiency by optimally employing existing technologies and inventing new process technologies. They also push their current production possibility frontiers continuously outward or invent new frontiers by introducing new products. While static efficiency is about optimality under given conditions, dynamic efficiency is about innovation with the objective to perform better in the future. According to [7.1], p.18 "… economic history and econometrics both suggest that the welfare gains from being dynamically efficient are much greater that those from being statically efficient." In relation to electricity generation capacity, statically efficient strategies will provide sufficient capacity based on current technology, while dynamically efficient strategies will give incentives to develop innovative solutions that are cheaper and/or result in higher customer satisfaction.

In the process leading to the National Electricity Market in Australia, one of the objectives was to encourage greater demand participation and to ensure competitive pressures to create a more efficient electricity industry, factors that can be summed up as dynamic efficiency. The issue of allocative (static) efficiency did not feature high in the discussions [7.13].

Invisibility

With invisible strategies, each market actor pursues his or her own objectives without worrying about anyone else's. Perfectly competitive markets provide overall production and consumption efficiency without anyone having to worry about attaining them. In principle, this should also keep costs at a low level, but examples from the introduction of electricity markets in California and the UK (retail access) suggest that this is not always the case. Invisible strategies also allow for a high degree of free choice, normally considered superior to strategies that involve direct command and coercion.

Robustness

The future is uncertain, which means that strategies will have to operate in environments for which they were not designed. Robust strategies operate reasonably well in such situations. Thus, a robust policy for peaking capacity will work reasonably well for example when demand appears to deviate from forecasts.

Timeliness

An important aspect is the ability of a strategy to be employed at the right time. For example in power systems with a current capacity shortage, it is necessary that a strategy can be effectuated immediately. On the other hand, in systems with a current surplus, there is more time to design long-term optimal policies.

Stakeholder equity

As stated in [7.1], "Gains and losses in efficiency are difficult to see, but policy winners and losers are highly visible". Consequently it is important for any chosen strategy that stakeholder equity is taken into account. It is well known that (statically) efficient market solutions by no means have to be equitable. An example is when prices in pure electricity spot markets become so high that weaker groups in society are adversely affected in an unacceptable way.

Corrigibility

As mentioned under robustness, unforeseen developments will always occur. In addition, policies will produce some unintended and possibly harmful results. Thus, it is important that a policy is corrigible.

Acceptability

Strategies that perform well on most criteria may still be politically or otherwise unacceptable in society. A relevant example is the introduction or increase of “green taxes”, designed to reduce the consumption of environmentally harmful products like energy. There are many good arguments for the introduction of such taxes, but it is often difficult to obtain enough political acceptability. Acceptability is related to stakeholder equity and simplicity. It is easier to introduce a policy that makes no stakeholders significantly worse off, and that is simple to understand for the parties involved.

Simplicity

Policies that are simple and understandable are clearly preferable above complicated strategies *ceteris paribus*. Too complicated policies will not work even when they are optimal from a theoretical point of view, because the market will not understand them. This is especially true for mass consumption products like electricity. Simple rules encourage participation, whereas complex rules are a barrier to entry. A resulting drawback of too complicated policies is that incumbents have a considerable advantage over new entrants based on their experience. Thus too much complexity creates a barrier to entry.

Cost

The cost of implementation is an important aspect of each policy. It must be taken into account when static efficiency is assessed. But even if a program is statically efficient when transactions costs are taken into account, it may still fail to satisfy a cost criterion if initial investment costs are very high, or if costs and benefits are unevenly distributed between stakeholders.

In the specific case of peaking capacity in an electricity market, the following criterion is also of great importance:

System security

The policy’s ability to obtain an acceptable level of system security.

7.2 Policy description

In this section, the major policies that are presently implemented in various systems are described. A discussion of each policy in relation to the criteria in the previous section is given in Section 7.3.

In [7.2], Jaffe and Felder argue that the benefit (increased system security) of marginal generation capacity (used for reserves only during a small number of peak demand hours), accrue largely to parties other than the owner of the capacity. This creates an externality, which according to economic theory will result in underprovision in a competitive market.

The authors discuss how to internalize this externality. They distinguish between strategies that fix a capacity price, and let the market find the amount, and strategies that fix the amount, and let the market find a price. They argue that one method is not inherently better than the other, but that the results depend especially on the forms of the marginal benefit and cost functions around the optimum. Setting the quantity at its conceived optimal level will lead to smaller errors than price setting, if the marginal benefit is declining rapidly at its intersection with marginal cost. In this case there is a kind of threshold, above which the marginal benefit of capacity is low, and below which it is high. The optimum must then be near this threshold. If on the other hand, the marginal benefit function is very flat in the vicinity of its intersection with marginal cost, the correct optimal level of capacity depends a lot on the marginal cost function. In such cases it would be better to fix the price. It is not evident that the marginal benefit function shows any of these characteristics, so there is no reason to prefer any policy *a priori* on these grounds.

7.2.1 Capacity Obligations

The objective of the capacity obligation policy is to ensure that the capacity levels necessary to maintain system reliability continue to be available after restructuring. This is a fixed capacity strategy. Similar capacity obligations existed in the traditional market structure, based on explicit reliability requirements, e.g.:

> "Each Area's resources will be planned in such a manner that, (...) the probability of disconnecting non-interruptible customers due to resource deficiencies, on the average, will be no more than once in ten years." [7.6].

As such, this policy comes closest to traditional methods.

A forecast for some planning period (e.g. year, month, day-ahead) is determined to establish the level of capacity resources that will provide an acceptable level of reliability consistent with agreed upon standards. Based on this forecast, a requirement is established to ensure a sufficient amount of capacity to meet the forecast load plus reserves to provide for unit unavailability, forecasting uncertainty and planned maintenance. Specific obligations are determined for “Load Serving Entities” according to some relevant criteria. As an example, the PJM Interconnection uses the following formula for the Accounted-For Obligation for each billing month during a planning period [7.3]:

$$\text{Accounted-For Obligation} = [(FSP \cdot DF) - ALM] \cdot FPR/100$$

where:

FSP - the summation of the weather-adjusted actual coincident summer peak for the previous summer of the end-users for which the Party was responsible on that billing day

DF - the Diversity Factor for the Zone (a factor calculated to share the benefits of load diversity between parties)

ALM - the active load management credits of the Party

FPR - the Forecast Pool Requirement, i.e. the amount, stated in percent, equal to one hundred plus the percent reserve margin for the PJM Control Area.

Similar formulas are used for the other time scales. Other examples are New England [7.4] and the New York [7.5]. A more advanced method for calculating participants' obligations is proposed in [7.7].

To comply with its capacity obligation, a participant or Load Serving Entity can either employ its own resources (including approved forms of load management) or buy capacity credits from other participants. Buying and selling of capacity credits between participants can in principal take place in bilateral or centrally organized markets. The PJM Interconnection operates Daily and Monthly Capacity Credit Markets for this purpose. The Capacity Credit Markets permit some participants to offer to sell and other participants to bid to buy credits to use generation capacity to meet their generation and capacity obligations. Other examples are the monthly market for Installed Capacity and the daily market for Operable Capacity in New England, which operated until March and June 2000 respectively.

The combination of setting a capacity obligation and creating a capacity market creates an identifiable commercial product associated with generation adequacy [7.9]. Because market prices may be very volatile (cf. Figure 7-1), these markets do not necessarily reduce the uncertainty for capacity investments. The reliability level is determined centrally, so there is no interaction between this level and the level actually desired by consumers.

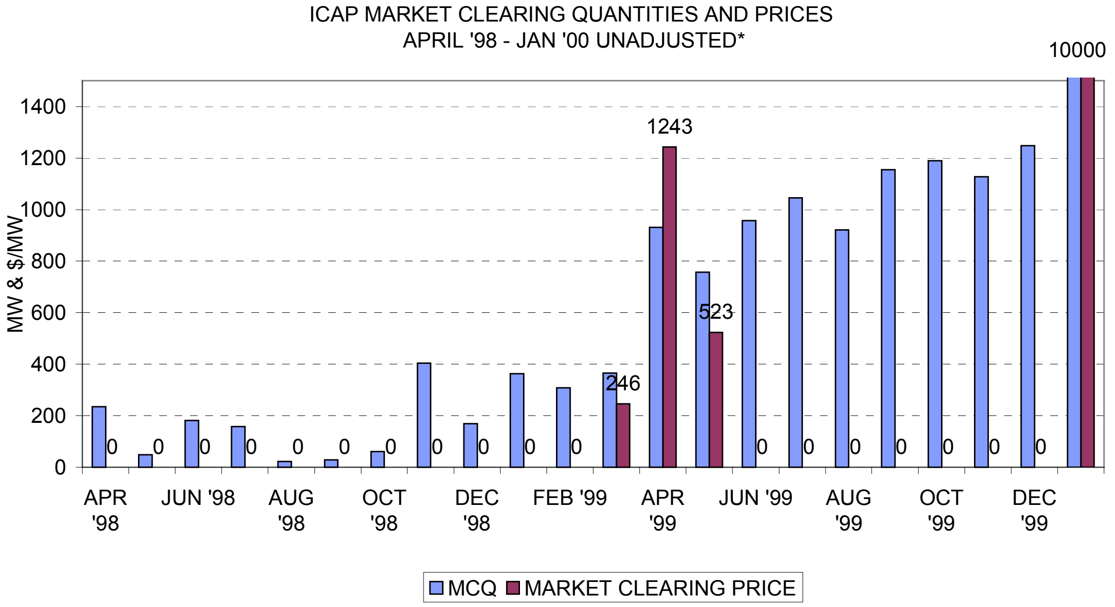

Figure 7-1: Quantities and prices in the montly capacity market in New Englang (source: ISO New England)

7.2.2 Fixed Capacity Payment

In a market with fixed capacity payments, an *ex ante* determined additional payment is made to all generation that is *available* during peaking hours. The objective of such payments is to distribute the income that generators would have in the small number of hours with very high prices throughout all the periods of operation. In the Jaffe and Felder framework, this is a fixed price strategy.

There are two principal motivations for this approach [7.9]: partly stabilization of generators' incomes, especially for peaking units and direct promotion of an extra level of generation adequacy by stimulating investments and discouraging early retirement. The fixed payments may be combined with a price cap because high prices should not be necessary when generators are partly remunerated for their investments. If the market is functioning well (no use of market power) and the capacity payment is appropriate, the price cap should normally not be necessary.

The main challenges for this method are to find the appropriate level of total remuneration, and its division between generators and possibly consumers that are willing to contribute with load shedding. In principle, total remuneration should compensate generators for costs that they do not recover from the pool. This is not an easy criterion to use in practice, because it is not possible to know future market prices, so it is not possible to forecast with certainty the level of cost recovery. Moreover, it is not evident how much of their cost generators actually should recover. If imprudent investments are made, these should obviously not be recovered, but it is not evident how to establish the level of investment costs that should reasonably be recovered. Given that the total level of remuneration is established, it should be distributed to participants according to their contribution to system reliability. One way is to use an "equivalent capacity" for each unit [7.10], equal to the capacity of a generator with availability 1.0 with the same contribution to reliability.

Fixed capacity payments are presently used in Spain, Argentina, Colombia and Chile. In Spain, the "Garantía de Potencia" or power guarantee was introduced as a central element in the deregulation process [7.11]. In 1998, the level of the payment was based on 70 % of the annual investment cost in a new combined cycle power station. Thus, a unit with these investment costs would recover 70 % of it investment costs if it were available during all the 4500 hours the payment is made. Payment is based on *past availability* and not on future guarantee. Some modifications were made in 1999, generally reducing the level of the payments somewhat. Consumers pay for the power guarantee through an uplift on the spot price.

7.2.3 Dynamic Capacity Payment

A system with dynamic capacity payments has been in operation in the Pool of England and Wales since 1990[1]. In this Pool, generators bid prices at which they will provide electricity throughout the following day. In addition, generators make availability declarations for each unit for each half-hour. The day-ahead prices and availability declarations are input to a unit commitment type optimization program to determine the minimum cost commitment and dispatch solution, where costs are based on the generators' bids. The System Marginal Price *SMP* for each half-hour of the following day is the price bid of the marginal generator.

The basic idea behind the dynamic capacity payment method is that the expected marginal cost of disruption should be reflected in the *ex ante* pool price. This is obtained by adding a capacity charge *CC* to the System Marginal Price *SMP*, which depends on the actual probability of loss of load, *LOLP*. The Pool Purchase Price *PPP* is calculated according to:

$$PPP = SMP + CC = SMP + LOLP \cdot (VOLL - SMP)$$

where *VOLL* is the Value of Lost Load, set by the Director General of the Office of Electricity Regulation (OFFER[2]), originally at GBP 2 per kWh and subsequently increased with the Retail Price Index. *LOLP* is calculated by NGC, based on the principles outlined in Section 2.2. Generators are paid for available capacity according to $LOLP \cdot (VOLL - \max(SMP, \text{bid price}))$, which is equal to *CC* for dispatched generators, and slightly less for others, depending on their bid price (which normally is higher than SMP, otherwise they would have been committed). Under normal, low load conditions *LOLP* will be very small, and *CC* is zero, cf. the discussion in Section 4.3.1. When system security is reduced due to reduced availability of capacity, *LOLP* increases, and *CC* increases. Under specific circumstances, *CC* can be the dominating ingredient of *PPP*. On an annual basis, the share of the capacity charge has varied considerably, but was between 10 and 20 % of the Pool Selling Price between 1995 and 1997.

The Pool Selling Price *PSP* is equal to *PPP* plus an Uplift, which compensates generators for reserve, plant available but not used and startup costs and ancillary services. Uplift is zero during low load ("Table B") periods, and is the only element of the spot price that is calculated *ex post*.

[1] A major restructuring of the England and Wales (in this connection often called "UK") market takes place in the autumn of 2000, where both the Pool and the capacity payment will be abolished, cf. [7.8]. The description here refers to the original solution.

[2] In 1999, OFFER was combined with Ofgas, the Office of Gas Supply to form OFGEM, the Office of Gas and Electricity Markets.

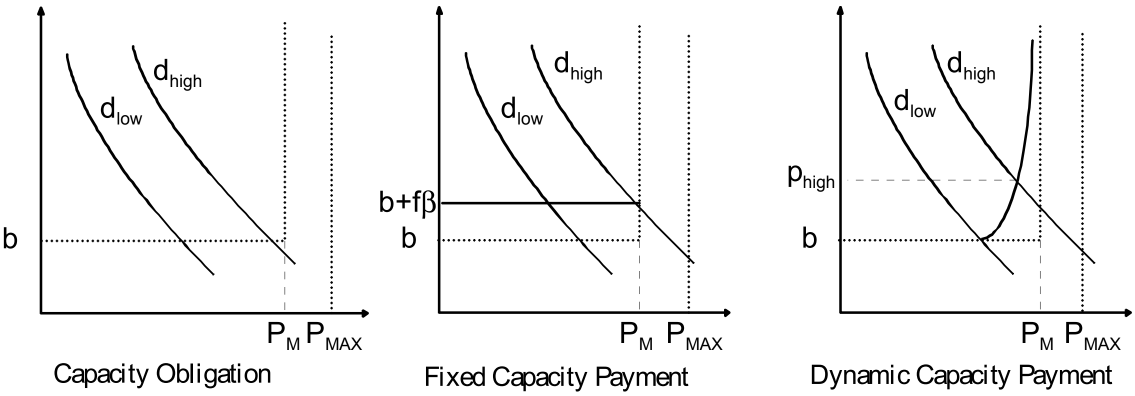

Figure 7-2: Illustration of Capacity Obligation and Fixed and Dynamic Capacity Payment

Figure 7-2 illustrates the concepts discussed so far. Each of the three panels shows two demand curves for high and low demand respectively. As in Chapter 3, *b* represents marginal generation cost and β annualized investment cost. P_{MAX} is available capacity, P_M available capacity minus reserves and *f* is a factor between 0 and 1.

With a capacity obligation, it is assumed that a capacity shortage will not occur, which means that available capacity P_M always exceeds demand. In this case, the price can always equal marginal generation cost. Generators are paid for fixed costs in the capacity market. In the fixed capacity payment method, some share of the investment cost β is added to marginal cost to remunerate capacity costs. The price should be set high enough to limit demand to available capacity, or viewed from the supply side, to guarantee sufficient capacity investment. In the dynamic capacity payment method, the price depends on actual reliability. Thus prices will be lower under low load conditions for the dynamic rather than for the fixed capacity method, and higher under high load conditions.

7.2.4 Energy-Only

In an energy-only market, there are no explicit provisions to ensure (long-term) generation adequacy. Prices are determined purely through the interaction of supply and demand, with minimal interference. Like in other electricity markets, provisions to guarantee short-term security are necessary in the form of ancillary service markets, but generation capacity *per se* receives no separate remuneration in the spot market.

While the other models spread the cost of supply in the peak-demand periods over many hours, in the energy-only model prices are only affected by this cost in the hours where marginal peaking capacity is price-setting.

If peak demand is supplied by some specific generation technology with low investment but high marginal cost, this unit can only generate a positive profit if price is above its marginal cost for a number of hours. This technology would have to set price at a level that covers it fixed costs during these hours[3]. If the market is in doubt if such prices can occur, investment will not take place. If demand is inelastic, involuntary load shedding will be the result.

The energy-only market requires least governance and reliance on non-market procedures, and consequently the possibilities for distortions are fewer than in other market structures. Energy-only markets are operating in Norway and California[4].

7.2.5 Proxy market prices

Other things being equal, it is natural to assume that the probability of capacity shortage is greatest in energy-only markets. Thus, especially in such markets it is imperative to design mechanisms to handle capacity shortages. One mechanism is to determine an administrative proxy market price to be used when the market does not clear. This may be viewed as an extension of the energy-only market, but it is treated as a separate policy, especially because of the extension proposed in [7.12], and explained below.

The proxy market price may be viewed in two ways: as a surrogate default bid, representing the price at which customers will be indifferent to having their loads curtailed, or as a price cap on the market price [7.14]. A purely surrogate-bid view focuses on ensuring that the price is set to a financially appropriate value (i.e. the best estimate of the value of lost load for the consumers concerned) during involuntary load shedding. There would be no consideration of whether load shedding could have been avoided by raising the price, or whether risk management considerations required it to be reduced. In a purely price-cap perspective, focus is on minimizing the need for intervention to achieve a desired level of reliability consistent with maintaining tolerable levels of economic risk for market participants. A too low price would result in too frequent market intervention, while a too high price might make it impossible to hedge risk, rendering the market inviable.

The surrogate-bid paradigm would be relevant if it was concluded that the market would not be capable of consistently clearing without intervention. It would help ensure a financially appropriate price was applied when the inevitable intervention occurred. If on the other hand, it was expected that the market would eventually be able to clear consistently and efficiently, provided it was given sufficient price freedom, the price cap paradigm becomes the more appropriate approach.

Administrative prices are used for handling capacity shortage in Sweden and in the National Electricity Market in Australia[5]. Sweden uses a price of SEK 3000 (USD 360) per

[3] With annual gas turbine costs of, say USD 40/kW and an expected annual number of 10 hours where the gas turbine is price setting, the price would have to be equal to USD 4/kWh, the order of magnitude of the Value of Lost Load in many systems.

[4] The Californian market has a price cap of USD 250/MWh, and consequently strictly belongs in the next section. However, the price cap in California is of a more general character, not primarily designed as a last resort to use when the market does not clear. The Norwegian market does not have any formally stated price caps.

MWh[6] in the regulating power market when there is a risk of shortage, and SEK 9000/MWh when consumers are actually disconnected. The National Energy Market in Australia presently uses a price of AUD 5000 (USD 3200) per MWh, but an escalation to AUD 20000/MWh by April 2002 has been proposed.

When a situation where a (day-ahead or real-time) market does not clear arises, there must be one or several market participants that do not have the contractual or physical resources to meet their obligations. They will bid in the market for their deficiency, and if there are insufficient supply offers, the market will not clear. For example, a trader may have a supply contract with a capacity limit, but an obligation to sell to general demand without such a limit (a common situation in the Norwegian market). Ideally, customers of this trader should be disconnected, and there would be a settlement between the trader and his customers that would not affect the other market participants. However, in a deregulated market, there will often be no direct connection between supplier and consumer. Combined with the fact that these relations are changing constantly, it is impossible to only disconnect consumers buying from the supplier causing the deficit. Thus, if load shedding is necessary, random consumers will be affected. The market participants causing the imbalance may be affected, but this is by no means guaranteed, depending on where and how load shedding is performed. The agents with a contractual deficit should be obliged to buy this deficit on the (day-ahead or real-time) market at the prevailing (administrative) price to expose them to the proper risk. Thus, as suggested in [7.12], the correct volume of disconnected demand should be estimated, and the responsible suppliers should pay the proxy market price to make up for their imbalance[7]. This is not an easy task, because demand that has been disconnected cannot be measured, but reasonable procedures to estimate demand can be established in the market rules. Ideally, the amount paid by the responsible suppliers should be used to compensate the consumers involved. With the appropriate level of the proxy price, this procedure should create the necessary incentives for all parties to ensure a positive balance also during extreme demand conditions.

Thus, a proxy market price is determined. This price balances the considerations between being economically representative in case of load shedding on the one hand and avoiding unacceptable financial risk for market participants on the other hand. When the market does not clear, the proxy price becomes the effective market price. All market participants that sell power on the relevant market receive this price. Participants that buy power pay this price. An estimate is made of the amount of load shedding, which is included in the balance for each market participant[8].

[5] The National Electricity Market in Australia also has a “reliability safety net” that gives the System Operator additional powers if system security is threatened.

[6] USD exchange rates in this Section reflect the December 1999 level.

[7] The resulting demand function is similar to the inelastic demand function introduced in Section 3.3.3.

[8] A complication occurs because of the need for reserves. All participants can have a non-negative energy balance, but there may be no capacity available for reserves. One way to solve this problem is to use an option-like market for reserves, similar to the one introduced in Norway in 2000. With this solution, the System Operator buys the right to use capacity for reserve purposes when this is deemed

7.3 Policy discussion

We are now in a position to compare the various policies from the previous section with respect to the criteria in Section 7.1. To facilitate the comparison, this is done on a criterion basis.

7.3.1 Static efficiency

Static efficiency is about maximizing welfare in the economic sense, expressed by the "society view" expression:

$$\underset{p_n, Q_m, q_{nm}}{\text{MAX}} \quad W =$$

$$\int_n \int_\theta \int_u \int_0^{q_n} (P_n(q',u,\theta)dq' - \sum_m C_n(q',u,m))dq'\, dFdGdn - \sum_m \beta_m Q_m \qquad \textbf{(7-1)}$$

subject to reliability constraints

The same symbol definitions apply as in Chapter 6. n is the time variable, θ the consumer type, u the stochastic variable and $q_n(p,u_n,\theta)$ demand of consumer θ in period n. $P_n(\cdot)$ and $C_n(\cdot)$ are consumer marginal willingness to pay and marginal cost of generation respectively. Q_m and β_m are the amount of installed capacity and the annual fixed cost of technology m, and q_{nm} is the generation in period n by technology m. p_n is the price in period n. A general multi-period multi-technology model is described in [7.15], but will not be pursued further here.

In the capacity obligation model, ΣQ_m is chosen based on a load forecast and reliability requirements. Most of the literature referred in Section 3.2 concludes that in such cases optimal prices in peak load periods must be in excess of marginal cost. However, if the market is competitive, prices will be driven down to marginal costs. The shadow price of the reliability constraint will never be positive, because the whole point of the model is to avoid this situation. Capacity costs are remunerated in the capacity market. Given the continued use of traditional methods, there is a significant chance that the traditional methods will continue to result in too high a level of total capacity.

In the fixed capacity payment model, an addition to marginal cost $f{\cdot}\beta$ is used under peak load, where $0 < f \leq 1$, which in principle is in accordance with the theory for a regulated system. However, in a deregulated system it is left to the market to determine Q_m, which may

necessary, e.g. on a seasonal basis. When extreme demand is expected, the System Operator will notify the market that the option will be exercised, and the sellers of these options, cannot use this capacity in the energy market. Thus, the proxy price mechanism will be active in the energy market, while there are sufficient reserves available. The analysis in Chapter 4 shows that it may be better to reduce the reserve requirements in such cases.

result in values under or over the optimal level if the fixed charge is not adapted or too late adapted to changing conditions.

In the dynamic capacity payment model, a variable addition to marginal cost *cc(u)* is used, which depends on the current status of the system. Although this resulting price is not a *market* determined price, it still contains a theoretically correct element regarding capacity shortage, if it is assumed that the calculation of *LOLP* is sound and that *VOLL* mirrors consumers' real losses of being disconnected. *cc(u)* depends both on demand and generation availability, and will generally depend on the margin between installed capacity and peak demand. This model is thus statically more efficient than the capacity obligation and fixed payment models.

The static efficiency of the energy-only model depends very much on the real elasticity of demand. If demand is elastic under all circumstances, the model actually represents a spot market, which is statically efficient[9]. If, on the other hand, demand is not very elastic in the short run, and only random rationing can be used to maintain system security, the model is not very efficient.

The proxy market price model addresses this point. In this case, prices in the case of market failure due to capacity shortage are clearly defined, and the market players can anticipate these prices. In the sense of the problem statement **(7-1)**, the problem is infeasible if a capacity shortage occurs, and is made feasible by introducing a proxy market price. To be statically efficient, the proxy price must reflect the real willingness to pay of the marginal disconnected consumer. It is clear that this is not possible, because this price will be situation dependent. Finally, rationed demand must be estimated, resulting in another source of inefficiency. Thus, this model is statically less efficient.

The capacity subscription model with optimal allocation of excess capacity has been shown statically efficient in Chapter 6. With random allocation, this strategy is not efficient. However, the degree of inefficiency is probably less in the latter two models than in the capacity obligation and payment models, because inefficiency only occurs when there is actually a shortage, normally a small number of hours.

7.3.2 Dynamic efficiency

In the capacity obligation model, suppliers have an incentive to introduce demand side participation, as long as this gives them a credit in view of their obligation (which it does explicitly e.g. in the PJM model). This is especially so because it will be very expensive to satisfy the obligation by investing in peak load capacity with a very short load factor.

In the fixed capacity payment model, there is little incentive to introduce innovative solutions. Suppliers are paid for available capacity anyway. In both these models, it may be difficult to define the demand side contribution exactly, which is important because this has a direct impact on the parties' economic result. This may reduce the degree of demand side

[9] Network effects, that potentially can have a negative effect on the spot market efficiency, are not taken into account in this argument.

participation. In neither of these models, the demand side is directly motivated to take innovative solutions into use, this is only done indirectly when approached by suppliers.

The dynamic capacity payment model may occasionally (depending on available capacity) result in high spot prices. If many consumers directly pay spot prices, they have an incentive to reduce demand in such periods. However, the model has no special properties to make consumers buy on the spot market. Consequently, neither of these three models has a strong dynamic efficiency with respect to involving the demand side.

In the energy-only model, consumers may be involuntary rationed. To avoid the negative effects of this, consumers with a high value of lost load may invest in backup power supply solutions. This may create some innovative development in this field, but it is difficult to see that this will be a great effect. However, in systems with an energy-only market, there will be a necessity to take measures to avoid involuntary rationing. If one wants to preserve the energy-only system, the proxy market price model is a logical next step. This model motivates all parties with a balance obligation to search for cost-effective solutions to ensure balance during peak-load conditions to avoid buying at the high administrative market price.

The capacity subscription model offers the greatest opportunity for innovative solutions, if implemented with this objective. This can be realized in the following way:

- A basic, simple and relatively cheap “fuse” device and communication solution is implemented for all consumers. The fuse, once activated, operates as a conventional fuse: it “blows”, reducing demand to zero. The consumer must manually reduce demand below his fuse size, before he can switch on the fuse again. The communication solution is also very simple: it only has to support a signal activating or deactivating the fuse.
- Consumers who find this basic solution unattractive can, as an alternative to buying more fuse capacity invest in a number of technological solutions that keep demand below fuse size in the way that best suits their preferences.

Thus, the capacity subscription model creates a demand for innovative solutions from the consumer side, in contrast to the other models.

7.3.3 Invisibility

Invisibility implies that market actors make their own, free choices, while market externalities are internalized in the market structure. Ideally, there are no externalities left – they have all been internalized.

The capacity obligation model is very “visible”: instead of leaving it to the market agents to decide how much to invest in generation capacity, a rigid control of their forecasted balance coerces each agent to have a positive balance.

The capacity payment models, the energy-only model and the proxy market price model are all invisible. This is typical for models relying on prices instead of volume control.

Also the capacity subscription model is invisible in this context: a market structure is defined, and it is up to the market agents to act within this structure. There is no control of volumes.

7.3.4 Robustness

From an engineering point of view, the capacity obligation model is probably most robust. It is more or less a continuation of methods from the past, which have proven to result in high levels of reliability. In this sense, the robustness of all other policies is more uncertain, because there is there is no long experience with any of them[10]. Clearly, when it comes to "engineering" robustness, control policies are more certain to satisfy their objectives than price-based solutions.

From an economic point of view, a robust solution will perform satisfactory in economic terms, even if conditions *ex post* turn out to be different from the *ex ante* anticipations. From this point of view, price-based policies are more robust, especially if prices can react on changing conditions. So in economic terms, the capacity obligation model is the least robust, followed by the fixed capacity payment model. The dynamic capacity model is more robust, because the capacity payment varies according to the circumstances. The energy-only model is robust if demand is sufficiently elastic. If not, it may not be very robust when available capacity becomes a binding constraint. The proxy market price model attempts to promote demand elasticity by using very high prices if there is a shortage. If this solution is to be robust, the proxy price must be adapted to the actual circumstances, which can be difficult. The capacity subscription model is robust in economic terms: capacity (fuse) prices will adjust to ensure balance between supply and demand capacity.

7.3.5 Timeliness

The criterion of timeliness is important when a capacity shortage is actually threatening in the short term, like Sweden.

The energy-only market is the easiest and quickest to implement – actually it will be the point of departure in many cases. Proxy market prices can also be implemented quickly, although some time is needed to build the structure and the systems to ration demand as effectively as possible[11], to estimate rationed demand ex post, to measure or estimate each market player's net obligations during a shortage etc.

[10] The dynamic capacity payment model has been used in England and Wales for 10 years, and there has been no capacity shortage. As discussed earlier, it is unclear if this is because or in spite of this model. The energy-only model has been in operation in Norway for 9 years, but because of the huge capacity surplus when the market was restructured, the Norwegian experience offers no evidence if an energy-only model is sustainable in the long run.

[11] Although random rationing is used, efforts should be made to ration as little demand as possible, while keeping system security at a satisfactory level. This is not a trivial task.

The capacity obligation and payment models need more overhead in the form of organization and software, and take some more time to implement. For the former model, additional time is required to allow the market participants to make appropriate arrangements, depending on the pre-restructuring situation.

The capacity subscription model requires most time to be implemented because all consumers must be involved to make the policy effective. This can be a long-term objective, while other policies are used in the short term.

7.3.6 Stakeholder equity

Before discussing stakeholder equity, it is necessary to define the relevant stakeholders, in this case:

- Small and medium-sized consumers
- Large industrial consumers
- Power producers
- System Operator

The difference between small and medium-sized consumers on the one hand and large industrial consumers on the other is that for the former electricity normally will constitute only a small share of their budget, while it is a substantial part for the latter group. Consequently, large industrial consumers have other conditions and a more explicit policy with regard to electricity purchases. The primary interest of the System Operator is to be able to operate the system within the reliability standards that have been defined. Although there are many more actors in the power market (e.g. distributors, traders, brokers), these are probably neutral as to what capacity model is chosen.

The next table summarizes these stakeholders' position with respect to the various policies. This is a very rough description, because the real effect will depend on the actual situation (surplus or deficit) and implementation details.

	small/medium consumers	large consumers	producers	System Operator
capacity obligation	-	-	+	+
fixed capacity payment	-	-	+	o
dynamic capacity payment	o	o	o	o
energy-only	o/-	o/-	o	-
proxy market price	+	+	o	+
capacity subscription	o	+	+	o

The capacity obligation model is attractive for producers, because it guarantees payment for capacity, even though this is market-based, which should result in low prices if there is excess capacity and the market is competitive. For consumers, it is negative, because it adds an extra cost element, which should not be necessary. It maintains the old situation where consumers preferences are decided on behalf of them, instead of letting consumers reveal their preferences. The system operator can build on the long experience that this model provides sufficient reliability.

The fixed capacity payment model is less attractive for consumers, depending on tariff and metering details. If the cost is transferred to consumers as part of a pure energy kind of tariff, this is an expensive solution for consumers, especially consumers with a high load factor. A form of demand charge at least gives consumers an opportunity to affect their capacity costs, but demand charges do not give very good price signals, as discussed earlier. For producers this is an attractive option, because it covers a substantial share of their investment in peaking capacity.

The dynamic capacity payment model roughly has the same effect as the fixed payment model. If it can be easily manipulated by producers, it is more attractive to them, but otherwise it is probably somewhat less attractive to producers, because is leaves more uncertainty. Because the charge is situation dependent, it may result in either higher or lower total payments than in the fixed case for consumers.

The situation of the system operator is uncertain for both capacity payment models, because it is uncertain to what degree this model really will bring enough capacity to the market place to ensure traditional reliability criteria.

The proxy market price model is attractive for most consumers because it reduces the probability of involuntary load shedding. The effect may be more ambiguous for large consumers buying part of their demand on the spot market. For producers, the model may be less attractive, because they can be exposed to high prices if they have imbalances in their contract portfolio. The model is attractive for the System Operator, because it provides the instruments that are necessary to use rationing to maintain system security.

The capacity subscription model is in principle attractive for consumers, because it allows them to tailor their purchases of electricity to their preferences for energy and quality of supply, cf. Section 3.4. Small consumers may perceive the increased complexity as negative, while this should not be a problem for large consumers. Also for producers it should be attractive, because it offers the opportunity to receive capacity payments. For the System Operator, the model is attractive from a system security point of view, but may be perceived as problematic due to increased complexity, at least in a transitional phase.

Finally, there may be a difference between real and perceived effects, especially for small consumers. For example may incidental rationing due to capacity shortage be perceived as extremely negative, while the positive effect of overall lower prices may not even be noticed. The same is true for capacity subscription: some will argue that consumers have to pay for peaking capacity that they previously got free. Although that is clearly not true, it may be difficult to convince consumers and politicians of the advantage: overall lower cost, which is difficult to prove in an easy-to-understand way.

7.3.7 Corrigibility

Most of the policies discussed are corrigible: for all payment or price-based policies, these prices can easily be adjusted[12], and it is relatively easy to change to a different policy. The capacity obligation model is somewhat less corrigible – if too much capacity is built, there is no way to reduce capacity economically *ex post*. In the longer run however, the level of obligation can be adjusted. The least corrigible model is the capacity subscription model. If this model should turn out to be infeasible after it has been implemented, the investments in communication and "fuse" technologies and software have been made, and are effectively sunk. These investments are much higher than for the other price/payment-based policies.

7.3.8 Acceptability

Acceptability of the various policies is probably related to simplicity and *perceived* stakeholder equity. Generally, it is very difficult to judge the policies on this criterion. The results are very dependent on implementation details and the ability to communicate a policy's advantageous properties to politicians and the public. In this respect it will be most acceptable to implement policies that result in few changes from status quo, like the capacity obligation and capacity payment models.

7.3.9 Simplicity

The simplest model is clearly the energy-only model. The fixed capacity payment model is also rather simple, though some complication occurs through the definition of availability and the way the payments are transferred from consumer to producer. The dynamic capacity payment model is more complicated, because it depends on rather complicated calculations of the Loss Of Load Probability. The proxy market price model is comparatively simple, but has some complicating elements that have been discussed. The capacity obligation model is rather complicated: forecasts must be made, obligation must be divided between parties, audits may be necessary etc. As an example, a considerable amount of paper is used to describe the requirements in the PJM and New England markets.

Clearly, the most complicated solution is the capacity subscription model. A number of organizational, hardware and software constructions have to be in place, and all parties in the market are involved, not at least all households.

[12] Even if price adjustment is easy and relatively cheap, too frequent adjustments can have a adverse effect on the policy itself. If the level of the fixed capacity payment is varied regularly and unpredictably, the market will no longer see it as a fixed payment.

7.3.10 Implementation cost

The cost of a policy depends on the starting point from which it is initiated. The natural starting point is the traditionally organized power sector, where small (domestic) consumers pay a fixed charge plus energy tariff, while commercial and industrial consumers often pay a form of demand charge in addition.

From this starting point, the energy only policy entails low costs, viewed from the capacity point of view. Existing metering equipment can be used like before and few costly measures have to be taken by the System Operator. However, this depends somewhat on the degree of sophistication of rationing policies, which are a necessary part of this strategy. The proxy market price model is slightly more expensive, because an organizational structure must be built around the administration of these prices. Total costs depend heavily on how the market agents hedge the risk of having to pay the high administrative price, e.g. by investing in generation equipment or by activating the demand side.

The capacity obligation and payment solutions bring with them a relatively high cost of administration. This is probably especially true for the obligation model: system forecasts must be made, and obligations must be defined, calculated, audited and verified

The highest initial cost is probably incurred by the capacity subscription model, because a controllable fuse and communication solution has to be installed for all consumers. Although relatively simple solutions can be used, and mass production will reduce prices, the total number of consumers is high and will require a high cost. It is difficult to use this model only for some sectors of consumers, because this creates inequality between consumers who have to pay for peak demand, and others who do not[13]. Still, in a transitional phase, a partial introduction could be acceptable. There could also be an exception for consumers with a (standard) fuse below a certain level.

7.3.11 System security

An important issue is the effectiveness of the various policies in maintaining an acceptable level of system security. This is a difficult question, because it is not altogether clear what an *acceptable* level of security should be. One possibility is to use the traditional criteria, based on “engineering judgement” (e.g. 0.1 days of loss of load due to capacity shortage per year), but many economists have argued that this is an unnecessary high level of reliability. So a rephrasing of the criterion could be a policy’s effectiveness to maintain an *agreed upon* level of system security. From a market economics point of view, the best result would be obtained if each consumer could select his or her optimal level of security, given a market price for energy *and* security. This would make the question of the level of security superfluous, because it would be the result of market forces.

The capacity obligation policy is probably best at maintaining agreed levels of security, because it controls the amount of capacity directly, and criteria are based on many years of

[13] There is precedence though for this model in the existing use of demand charges for commercial customers only.

experience. The fixed capacity payment policy can also obtain a high level of security, if only the payment is set high enough, but there is a risk of too high a level, even after traditional criteria. The effect of the dynamic capacity payment model is more uncertain, which is also the case for the proxy market price model. While the former probably will not result in too high levels of security (because payments will become very low when such a situation is approaching), the latter may have this result, because the threat of extremely high prices may motivate overinvestment, even if the probability is small. In the energy only model, there is no way to control the level of installed capacity, and therefore, the level of security.

Finally, the capacity subscription model comes close to the ideal situation, where consumers can select their desired level of security by buying fuse capacity. So even if this model may not obtain the *traditional* level of security, it comes much closer to an *optimal* level. At the same time, the model legitimizes the use of (fuse based) rationing, which is barely if at all acceptable in the energy only model. This is an example of how technology progress (in metering and communications) can change the originally public good of security into a private good, cf. Section 3.2.3.

7.3.12 Policy analysis summary

Table 7-1 on the next page summarizes the discussions from Sections 7.3.1 to 7.3.11, as far as this can be done with a single good (+), satisfactory (o), poor (-) indicator.

Table 7-1: Summary policy evaluation

	obligation	fixed payment	dynamic payment	energy only	proxy prices	capacity subscription
static efficiency	-	-	o	-	o	+
dynamic efficiency	-	-	-	-	o	+
invisibility	-	o	o	+	o	+
robustness						
engineering	+	o	o	-	-	+
economic	-	-	o	-	o	+
timeliness	o	o	o	+	+	-
stakeholder equity	-	-	+	o	o	0
corrigibility	o	o	+	+	+	-
acceptability	+	+	o	o	o	-
simplicity	-	+	o	+	o	-
cost	o	o	o	+	+	-
system security	+	o	o	-	o	+

As could be expected, and was clear from the previous discussions, there is no clear best policy. Policies that perform well on some criteria perform poorly on other. However, by analyzing the results systematically according to various criteria, some tendencies become clear.

The simplest way to analyse the results numerically, is to assign the characteristics +, - and o numeral values of 1, -1 and 0 respectively, and adding the values for each policy. This gives the following result:

Table 7-2: Policy ranking; simple summation of criteria

	obligation	fixed payment	dynamic payment	energy only	proxy prices	capacity subscription
points	-3	-3	0	1	4	0
ranking	5	5	3	2	1	3

Alternatively, one can look at which policy performs best on most criteria, or worst on least:

Table 7-3: Policy ranking; best on number of criteria

	obligation	fixed payment	dynamic payment	energy only	proxy prices	capacity subscription
points	3	2	2	5	4	5
ranking	4	5	5	1	3	1

Table 7-4: Policy ranking; worst on number of criteria

	obligation	fixed payment	dynamic payment	energy only	proxy prices	capacity subscription
points	-6	-5	-2	-4	0	-5
ranking	6	4	2	3	1	4

This analysis assumes equal weights on the criteria. This may not be realistic, but on the other hand, it is not possible to give generally valid weights. If there already is a capacity deficiency, there is little time and it is natural to resort to methods that ensure reliability in the short run, i.e. the system security criterion is given much weight. On the other hand, in a system with ample capacity, static or dynamic efficiency will be given more weight.

The obligation and fixed payment policies are the two worst on all three simple ranking methods. Specifically, these methods perform badly on market efficiency criteria – essentially they are not market-based policies[14]. The proxy prices policy is among the two best on all ranking methods. The dynamic payment is "middle of the road", while energy only and capacity subscription are both best and worst, depending on the ranking method. However, the energy only method is disqualified, because it does not satisfy the system security requirement.

The proxy prices policy is a reasonable policy on most criteria. It is relatively easy, cheap and quick to implement. Because there is little experience with the method so far, there is some uncertainty with respect to if it is effective. One can anticipate, that the threat of having to buy power at rationing prices will motivate suppliers to avoid coming in a buying position in such cases, and that this will stimulate the adaptation of innovative solutions, especially on the demand side.

The capacity subscription policy looks very promising on the issues of efficiency, robustness and system security, but there is a considerable threshold to the introduction of this policy, which is caused by costs and complexity.

Concluding, in an early stage after restructuring it may be appropriate to resort to the capacity obligation or payment method if the capacity balance is tight at the time of transition. For the medium term, or if there is ample capacity initially, it is sensible to introduce proxy market prices to transfer the risk of a capacity deficit to market participants, with due attention to the appropriate price level. If there are signs that proxy market prices are not able to guarantee reliability, a long time objective can be to adapt capacity subscription, possibly building on an infrastructure already initiated, e.g. at the distribution level. The main reasons for moving towards capacity subscription are the efficiency gains, that are not the least caused by the possibility of consumers selecting their own levels of reliability.

7.4 References

[7.1] Harvey Averch, "Private Markets and Public Intervention", University of Pittsburgh Press, 1990.

[7.2] Adam B. Jaffe, Frank A. Felder, "Should Electricity Markets Have a Capacity Requirement? If So, How Should It Be Priced?", *The Electricity Journal*, December 1996, pp. 52-60.

[7.3] "Reliability Assurance Agreement among Load Serving Entities in the PJM Control Area", Schedule 7, updated September 19, 2000 [cited 2000-10-30]. Available from Internet <http://www.pjm.com/index.html>

[14] This view is supported in [7.16] for the obligation model: "The capacity markets are a holdover from the regulated setting, when capacity decisions were not made in response to price expectations. (...) once competitive electricity markets are established in New England, it would be appropriate for the capacity markets to terminate."

[7.4] "Composite Restated New England Power Pool Agreement", Forty-Second Amendment, Attachment 2, Section 12 [cited 2000-10-30]. Available from Internet <http://www.iso-ne.com/main.html>

[7.5] "Final Installed Capacity Manual", posted 3/31/2000 [cited 2000-10-30]. Available from Internet <http://www.nyiso.com/markets/>

[7.6] "Basic Criteria for Design and Operation of Interconnected Power Systems", Northeast Power Coordinating Council, Document A-2, August 9, 1995 [cited 2000-10-30]. Available from Internet <http://www.npcc.org/>

[7.7] Narayan S. Rau, "Assignment of Capability Obligation to Entities in Competitive Markets – The Concept of Reliability Equity", *IEEE Transaction on Power Systems*, Vol. 14, Nr. 3, pp. 884-889, August 1999.

[7.8] Richard Green, "Draining the Pool: the reform of electricity trading in England and Wales", *Energy Policy*, Vol. 27, 1999, pp 515-525.

[7.9] I.J. Pérez-Arriaga, "Reliability and Generation Adequacy", in "Reliability in the New Market Structure, Part 1", *IEEE Power Engineering Review*, Vol. 19, Nr. 12, pp. 4-14, December 1999.

[7.10] Juan J. Alba, "The Spanish electricity pool", EES-EUTP Electricity Markets, Fundamentals and International Experiences, Porto, Portugal, 26-30 October, 1998.

[7.11] J.L. Fernández González, G. González Morales, "Analysis of long term supply guarantee in the Spanish power system. Capacity payment and alternatives", DistribuTech Europe, Madrid, Spain, 28-30 September, 1999.

[7.12] Lennart Söder, "Who should be responsible for generation capacity addition", International Conference on Electric Utility Deregulation and Restructuring and Power Technologies 2000, London, England, 4-7 April 2000.

[7.13] Hugh Gleeson, "A Retailer's Perspective Of Australia's National Electricity Market - Where it has come from and where it is going", presented at CIGRE Study Committee 38 Colloquium "Experience with Electricity Markets", Perth, Australia, 29 September 1999.

[7.14] "Review of VoLL in the national electricity market", Issues Paper from the Australian National Electricity Code Administrator of 12 May 1999 [cited 1999-12-08]. Available from Internet <http://www.neca.com.au/>

[7.15] Michael A. Crew, Paul R. Kleindorfer, "Public Utility Economics", The Macmillan Press Ltd., London, 1979

[7.16] Peter Cramton, Robert Wilson, "A Review of ISO New England's Proposed Market Rules", ISO FERC Filing On The Market, NE Markets Review, September 9, 1998 [cited 2999-12-08]. Available from Internet <http://www.iso-ne.com/main.html>

Chapter 8: Conclusions

The point of departure for this thesis has been a restructured power system with an energy-only market solution – i.e. a market solution without explicit mechanisms to promote investment in peaking capacity. The focus of the work is on the mature power systems in highly industrialized countries with slow growth. Based on power system reliability theory on the one hand, and the effects of uncertainty and risk aversion on the other, it has been shown that in an energy-only market, it is impossible to guarantee the availability of the traditional level of generation capacity. To some extent, most obviously in Sweden, this is confirmed empirically.

I.	In a restructured power system with an energy-only market solution, it is not possible to guarantee traditional levels of system security.

Because it is not possible to guarantee the level of security, rationing cannot be ruled out. Traditional pricing theory also shows that optimal solutions do have a certain level of rationing. Random rationing is expensive and socially unacceptable.

II.	Some form of rationing during peaking conditions is unavoidable in an energy-only market. To make this acceptable and efficient, it must be voluntary and price based.

Consumers have different preferences for the use of electricity and its reliability. This is confirmed by surveys where consumers are asked to value uninterrupted supply. Traditional systems do not normally consider these differences. This is partly due to technological reasons: until recently, it has been difficult or at least very expensive to differentiate reliability. However, probably there has also been too much emphasis on a “one-size-fits-all” supply side oriented attitude.

III.	A more efficient electricity market will be obtained if differences in consumer preferences with respect to reliability are taken into account.

An important role of peaking capacity is to act as reserve capacity to support system security. Theoretically, the level of security in a power system should be such that the marginal cost of increasing this level equals the marginal benefit from lower outage costs. Due to the complexity of the system and the uncertainty with respect to both costs and benefits, this theoretically sound principle is not used as a practical guideline. Over time, a number of criteria based on engineering judgement have developed, e.g. a Loss of Load Probability due to generation capacity deficiency of less than 0.1 days per year. These criteria are met by requiring that certain levels of reserves are maintained in power systems. However, there is no one-to-one relationship between reserve levels and system security. In many situations, the required level of security could be obtained with lower reserve levels, especially if demand can be controlled quickly and efficiently.

IV.	By actively taking user preferences into account with respect to the security of supply and using efficient solutions for load control, the need for generation capacity is reduced. However, sufficient generation capacity must be kept available to avoid system collapse.

V.	Reserve requirements are an indirect way to ensure system reliability. The need for reserves and, consequently, peaking capacity can be reduced by targeting reliability measures like the Loss of Load Probability *LOLP* or Expected Energy not Served *EENS* directly. Research should be directed at the development of tools and methods for maintaining acceptable reliability without the use of fixed reserve levels.

It is reasonable that generators are paid for the provision of ancillary services (“activities that pertain to the provision of all electric services necessary for efficient and reliable generation, transmission and delivery of active power”). Efficient resource utilization suggests market based solutions insofar as market power problems can be avoided. In this thesis, a solution with explicit markets for three ancillary services is presented, where the markets for energy and the ancillary services clear simultaneously. Test calculations show that a market for ancillary services increases generator revenues and thus the profitability of investments in generator capacity. However, the increase is offset by a certain reduction in the energy price.

VI.	Markets for ancillary services are a tool for the efficient provision of these services. They provide some increased revenue for generators, but this is probably not enough to secure sufficient investment in generation capacity

Peaking capacity is needed to ensure uninterrupted supply during infrequent occasions of high demand. The kernel of the capacity balance problem in a restructured market is that prices do not reflect short-term market conditions. Consequently, consumers will not reduce their demand regardless of the spot price level, and random rationing becomes the only solution to reduce demand. An innovative solution to the problem is the introduction of capacity subscription. Each consumer subscribes on the level of capacity he/she prefers to use *during system peak conditions*, by buying a “fuse” limiting demand to this level. The “fuses” are activated by the System Operator *only* when unconstrained demand would exceed available capacity. Capacity is traded on a capacity market, and prices will reflect market conditions. The interesting features of this solution are that the demand for capacity will reflect consumers’ real preferences for reliability during peaking conditions, and that the consumer can weigh the procurement of capacity against advanced load control devices. Thus, this solution promotes the demand-driven development of innovative solutions.

VII.	Capacity subscription • guarantees that demand + reserve requirements will not exceed generation during peak conditions • is the only way to create a market for capacity, where consumer preferences for reliability together with the cost of supply directly determine the market price for capacity • promotes the development of innovative solutions for load control

A number of policies have been implemented around the world to ensure power system reliability through the availability of sufficient generation capacity. A comparison was made between the following:

- capacity obligation
- fixed capacity payment
- dynamic capacity payment
- energy-only
- proxy market price
- capacity subscription

The energy-only method was included for reference. It is not an acceptable policy, because it does not satisfy the requirement of system security.

No policy is optimal on all criteria. The obligation and fixed payment policies perform badly on market efficiency criteria – essentially, they are not market-based policies. The dynamic capacity payment method has some attractive properties, but depends on a pool-based market solution. The calculation of the payment is opaque, and it is probably prone to market power problems.

The proxy prices policy has many attractive properties. It is cheap and can be implemented quickly. Because there is little experience with the method so far, there is some uncertainty with respect to whether it is effective. A condition for this is that market participants with a deficiency actually are made to pay for this deficiency in the case of rationing.

The capacity subscription policy looks very promising on the issues of efficiency, robustness and system security, but there is a considerable threshold to the introduction of this policy, which is caused by costs and complexity.

VIII.	In an early stage after restructuring, it may be appropriate to resort to the capacity obligation or payment method if the capacity balance is tight at the time of transition. For the medium term, or if there is ample capacity initially, it is sensible to introduce proxy market prices to transfer the risk of a capacity deficit to market participants, with due attention to the appropriate price level. Capacity subscription can be the long-term objective. The main reasons for moving towards capacity subscription are the efficiency gains that are to a significant degree caused by the possibility of consumers selecting their own levels of reliability.