EPRI | ELECTRIC POWER RESEARCH INSTITUTE

A Framework for Valuing Demand Response as a Capacity Adequacy Resource

1017876

A Framework for Valuing Demand Response as a Capacity Adequacy Resource

1017876

Technical Update, December 2009

EPRI Project Manager

B. Neenan

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

ORGANIZATION(S) THAT PREPARED THIS DOCUMENT

Cornell University

EPRI

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

CITATIONS

This document was prepared by

Department of Applied Economics and Management
Cornell University
Ithaca, NY 14853

Principal Investigator
R. Boisvert

This document describes research sponsored by the Electric Power Research Institute (EPRI).

This publication is a corporate document that should be cited in the literature in the following manner:

A Framework for Valuing Demand Response as a Capacity Adequacy Resource. EPRI, Palo Alto, CA: 2009. 1017876.

PRODUCT DESCRIPTION

This report demonstrates the importance of customer participation in decisions about how much reliability, in the form of capacity adequacy, to provide electricity consumers in centralized organized markets operated by independent system operators/regional transmission organizations (ISO/RTOs) or by electric utilities.

Results and Findings

In the absence of consumers' affirmation of their willingness to pay for reliability, either too much or too little capacity will be provided or societal resources will be misallocated. Allowing consumers to designate how much capacity they want or, alternatively, treating specified load management capabilities (demand response) as supply resources can be equally effective in achieving the required degree of consumer participation.

Challenges and Objectives

The report will help those responsible for developing demand response programs to understand the underlying market mechanisms that determine why and to what extent consumers are willing to participate in load curtailment programs designed to reduce capacity costs, and how their participation affects capacity costs.

Applications, Value, and Use

Demand response programs are offered in some markets by utilities as part of retail service offerings and in other markets by ISO/RTO wholesale market operators. There is considerable discussion about the need for these programs, about which entities should provide them if they are appropriate, and about how much consumers should be paid to participate. The findings of this report will clear up many ambiguities and misconceptions that stymie the resolution of these issues and will focus attention on the fundamental question of what constitutes a robustly efficient and effective electricity market.

EPRI Perspective

EPRI seeks to help market designers and stakeholders understand the character and impacts of demand response, particularly as it relates to how consumer participation in electric markets influences market efficiency.

Approach

The approach of the report is to provide a conceptual framework for understanding the role and value of demand response in determining the level of service reliability that consumers are provided.

Keywords
Demand response
The value of demand response
Capacity market design
Demand subscription service
Load as a capacity resource

CONTENTS

1
INTRODUCTION

Promoting energy efficiency and demand and price response in electricity markets dates back at least to the late 1970s and early 1980s. The initial emphasis was on affecting consumption to reduce utility supply costs. Energy efficiency programs sought primarily to reduce the energy required to meet consumer and business needs – resource efficiency in the sense of achieving the same economic output and consumer satisfaction for fewer kWhs. To some degree, efficiency measures adopted also reduced peak demand, which lowers utility capital costs, achieving another form of economic efficiency. Demand response programs, primarily through utility imposed or directed load curtailments, were designed specifically to reduce peak capacity requirements, but often they also were used to resolve supply shortfalls.

The restructuring of electricity markets, which began in the late 1990s, resulted in the creation of centralized entities that are responsible for ensuring the reliability of the electric system under their aegis. The seven chartered ISO/RTOs are responsible for over 80% of the electricity supply in the United States. The growth of demand response in the United States, from around 3 GW prior to 2000 to over 3.7 GW in 2009 is in large part due to the increasing recognition that such programs are important to the efficient functioning of electricity markets (Goldman, *et al*. 2007). There is growing sentiment that these programs (particularly those in demand and price response) will help to maintain reliable and affordable electric service and to promote the efficient use of energy resources (Rohmund 2009).

This widespread interest in demand response is echoed in the Energy Independence and Security Act of 2007 (EISA 2007) which mandated that FERC, the Federal Energy Regulatory Commission, provide a projection of demand response potential. A recent report issued by FERC suggests that demand response could provide almost 20% of the country's capacity requirements by 2020 (FERC 2009). FERC was also asked to estimate how much of the potential could be achieved within five and 10-year time horizons, including options for funding and/or incentives for the development of demand response, to identify barriers to demand response programs, and to make recommendations for overcoming any barriers. While FERC did not so state, an implication of its findings are that a key to fostering demand response is to establish a consistent and credible means for ascertaining its value across diverse market conditions, and over time

The optimistic projections for reductions in peak load found in FERC's report to Congress depend on widespread investment in and the installation of Advanced Metering Infrastructure (AMI) by electric utilities throughout the country. Smart metering technology serves an enabling role when combined with other initiatives, such as the implementation of demand response programs, revised outage restoration practices, and the adoption of devices that communicate consumption and price/event information to consumers and the utility. As the adoption of AMI technology expands, private manufactures will also enjoy opportunities to produce a new generation of appliances and other electrical equipment that contain built-in control devices that facilitate demand response through direct communication with AMI.

A recent report by the IRC, comprised of the 10 Independent System Operators and Regional Transmission Organizations (ISO/RTOs) in North America, which serve about two-thirds of electricity consumers in the United States and more than 50 percent of those in Canada, shed light on the character of demand response today. According to their data for the United States, there are about 25 GW of curtailable demand response resource enrolled in various programs.[1] About 73% of this load is enrolled in capacity programs, while 13% and 11% are enrolled in ancillary service and energy-price programs, respectively. Only about 3% are enrolled in energy-voluntary programs, and these are only available in PJM and in the NYISO.

On average, only about 4% of the system peaks in these ISO/RTO operated markets are enrolled in demand response programs; the amount ranges from a high of about 7.5 % in MISO, to a low of about 2.5% in PJM, according to the 2007 IRC census. The amount of demand response available in the ISO-NE and PJM market has increased dramatically since then as the result of the commencement of a centralized capacity market that provides more and better opportunities for participation by loads: more about that later.

The enthusiasm and optimistic assessment of the potential for short-term demand response programs, including real time pricing (RTP) service in regulated and competitive markets and load curtailment programs offered by utilities and RTOs/ISOs, is not supported by a rigorous exposition of how demand response provides value. There is a paucity of research into demonstrating how much demand response is optimal either from the electricity market perspective (efficiency and reliability improvements), from the perspective of the electrical appliance industry contemplating a new generation of "smart" appliances, or from a public policy perspective. The most comprehensive review of how demand response is valued is in a report to Congress by the U.S. Department of Energy (2006).

There less than full agreement in the literature as to exactly how demand response programs contribute to market efficiency and reliability, the management of market risk, and the overall social welfare in regulated or restructured electricity markets. That is true both in terms of the mechanisms by which demand response delivery benefits, the level of those benefits, and to whom they accrue.

The critical issue is whether the benefits (e.g. the value of demand response resources), however distributed, are sufficient to justify the cost associated with their acquisition. The establishment of this value is of particular importance to decisions about the nature and extent of investment in Smart Metering, the justification for which in many cases depends on the how the technology enables greater demand response and associated benefits (Neenan and Hemphill 2008). Similarly, the justification for Smart Grid investments depends on the extent to which demand response is enabled to a greater degree.

This current situation is not all due to a lack of effort. To date, analysts have used a variety of methods to address this issue, and they have generated a wide range of results that are not easily compared or reconciled (U.S. Department of Energy 2006). The range in results arises in some cases because of fundamental differences in the market design, and in other cases they are due to market circumstances. The choice of methodologies in conducting the evaluations also can affect the results dramatically.

---

[1] The FERC estimate of 37 GW includes the demands response resources cited by the IRC plus those of utilities located in other areas not covered by an ISO/RTO.

The objective of research on which this report is based is to develop a comprehensive and rational framework to describe the process by which benefits can be attributed to demand response programs, specifically those associated with providing system reliability. This is accomplished by focusing on defining how customers value reliability, and from that explaining how the market price of capacity that provides reliability is determined. Initially, we begin with those programs that contribute to the system adequacy component of system reliability.[2]

System adequacy refers to planning to provide reliability in the long term, to which achieving short-term system security is inextricably linked. The framework developed to characterize demand response value is focused initially on competitive electricity markets operated by ISO/RTOs, but the results are extended to demand response resources in vertically integrated markets operated by regulated utilities. This comparison allows ascertaining the extent to which there may be a gap between the private and public value of demand response in regulated or ISO-run electricity markets.

The rest of the report is organized into nine sections. Section 2 provides a categorization of demand response programs that distinguishes those that involve dynamic pricing service from those that treat load management capabilities as a system resource, which is the concern of this report. Section 3 provides an overview of the economics of demand response, which provides a structure upon which a framework for valuing demand response as a capacity resource is constructed. That framework is developed in Sections 4-7, beginning with a definition of system reliability as it relates to electricity supply (Section 4), and then portraying how consumer demand for reliability should direct how it is supplied (Section 5). Sections 6 and 7 extend this framework to how capacity is procured in centralized wholesale ISO/RTO markets and by those managed by vertically integrated utilities, respectively, and the role of demand response in each. Section 8 suggests how the framework can be made more useful by developing an empirical analog.. Sections 9 and 10 contain supporting technical materials.

[2] As argued by Hemphill and Neenan (2008), the initial focus on capacity programs is justified in part because some the ISOs/RTOs, offer consumers opportunities to supply capacity to meet the market's capacity obligations from which we now have the benefits of several years of experience to form expectations about both the level of exposure to curtailment and the size of impacts from participation. The emergence of a group of curtailment service providers has also served to help customers evaluate participation and in some cases offer support services such as assistance in developing and executing load curtailment actions when events are declared.

2
THE NATURE OF DEMAND RESPONSE PROGRAMS

The objective of this research is to develop a framework to establish the value of demand response both as a source of capacity in electricity markets. To do so, it is critical to recognize at the outset that all capacity-focused demand response programs are not created equal. The several types of programs that are currently in place, or are being anticipated, differ both in their purpose and in essential design features. Those designed to allow participation in centrally managed capacity markets must be consistent with market protocols in order to ensure that there no adverse effect on the overall market design. Other programs that are run by vertically integrated utilities in states where electricity markets have not been restructured must reflect the character of the utility's resource planning processes, and be consistent with regulatory rules and regulations.

Although similar economic principles apply, the explicit recognition of these differences provides a level-playing field from which to understand the differences in the values of these customer-supplied resources. So, before developing a framework for valuing demand response as a capacity resource, it is first necessary to differentiate the ways in which customer load management can contribute to efficiency in electricity markets.

Classifying DR Programs

Demand response can be defined generally as the changes in electricity usage by end-use customers in response to changes in the price of electricity over time, or to highly focused incentive payments designed to induce lower electricity use at times of high market prices or when system reliability is jeopardized. The way in which changes in load influence market operations, and presumptively generate value to consumers, depends on the nature of the inducement.

Figure 2-1 differentiates demand response into two major categories: changes in load that result for customers acting in their own interests by changing usage based on the prices they face (the grouping on the right side of the diagram), and changes in load that result from customer responding to opportunities or obligations to change their usage due to a directive from the system operator, the two grouping on the left side of the diagram..

Non-dispatchable pricing plans provide consumers with a price schedule and they can buy as much as they want under the terms of that schedule. Consumers are price takers, and they exercise their economic power of choice by buying up to the point where the price of the good equals its marginal value of consumption. In some cases, the schedule offers a fixed price for all consumption at any time. Time-of-use schedules differentiate the usage price by the time of day. Others incorporate a dynamic aspect; prices are posted frequently, a day ahead or even for each hour. In every case, the supplier quotes firm, non-recourse prices and consumers decide how much to buy.

Dispatchable pricing plans have two structures. Economic plans allow consumers to alter their consumptions based on prevailing prices. A common form allows consumers to bid load

curtailments into the wholesale market and when called upon to do so, they curtail usage and are paid the market prices (usually the same price that generators are paid).

Under reliability plans customers contract to curtail and when called upon are obligated to do so or to pay a penalty if they fail to curtail. Alternatively, the curtailment may be enforced automatically through the activation of switches that shut off designated devices for some period of time. Dispatchable pricing plans share a common characteristic; the participating consumer has agreed to allow some entity, the system operator, its utility, or another market actor to determine when to activate the consumer's demand response, in most cases to ensure system reliability. To understand why these programs have been implemented, it is essential to explain the concept of reliability as it relates to electricity supply.

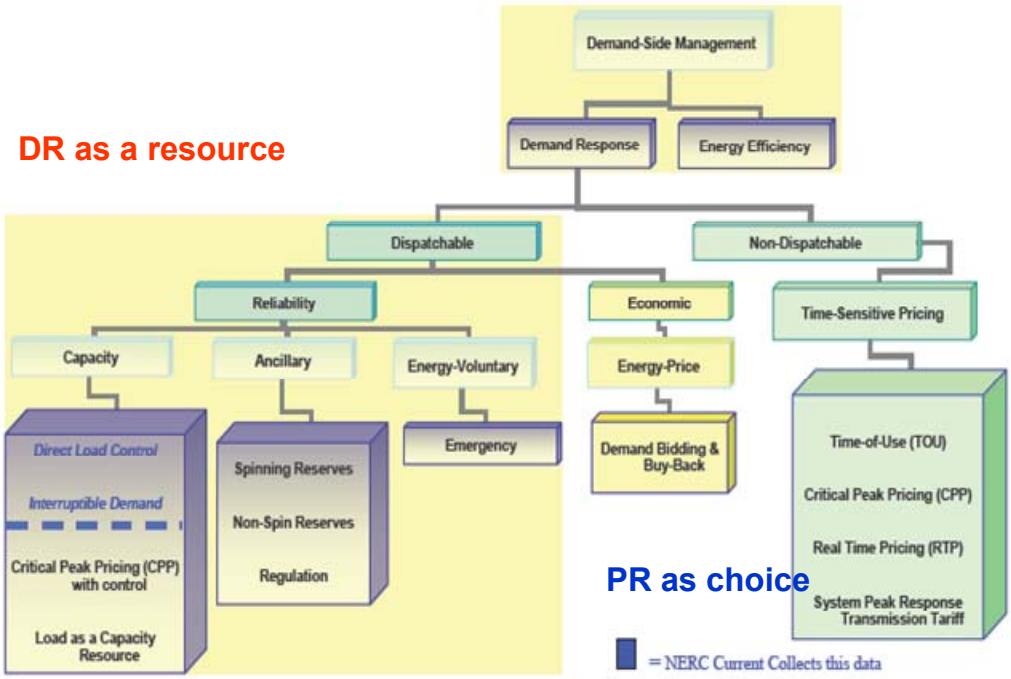
Electric System Reliability and Security

Electric systems are designed to provide consumers with electricity when and where they want to use it. They are comprised of generation units that produce electrical energy (the kWhs consumers use), high-voltage transmission systems that deliver the electrical energy to regional load centers, and the local distribution systems that lower the voltage to levels compatible with the requirements of consumer loads. Because electricity can not be stored economically to any substantial extent, and it plays such a vital role in our lives, the system is designed to achieve a very high degree of reliability, which is defined by the extent and/or frequency with which service is interrupted.

Within the context of system *reliability*, system resource adequacy relates to the long-term ability of the system to meet demand, under the explicit consideration of the inherent uncertainty and fluctuations in demand and supply, the non-storable nature of electricity, and the lead time needed to expand generation capacity. Adequacy is provided by building or acquiring sufficient generation and transmission assets so that peak demand can be achieved with a specified degree of certainty.

In contrast, system *security* relates to the short-term operational aspects of the system and its ability to withstand disturbances or contingencies. Security is provided by means of protection devices and operation standards and procedures that include security constrained dispatch and the requirement for so called ancillary services such as: balancing-up energy for voltage support, regulation, and spinning reserves.

The meeting of adequacy requirements and achieving security goals up to the early 1970s was accomplished almost completely by building and operating generating plants. At that time, utilities began to experiment with using consumers' load management capabilities to reduce the generation plant requirements both for adequacy and security purposes. The focus of this study is on how the level of system adequacy is determined and fulfilled and the role and value of demand response in meeting system adequacy requirements.



Source: North American Electric Reliability Corporation (NERC). December 2007. Data Collection for Demand-Side Management for Quantifying its Influence on Reliability, Results and Recommendations. Princeton, NJ. NERC_DSMTF_Report_040308.pdf.

Figure 2-1

3
ECONOMICS OF DEMAND RESPONSE

The objective of this report is to provide insight into how demand response provides benefits through changes in how resource adequacy requirements are achieved. Characterizing both the level and distribution of benefits is essential because market designers, policy makers, regulators and consumers have a shared interest in understanding who gains and who loses and by how much. Moreover, a framework to value demand response as a capacity resource must, at a minimum, be able to accomplish two things. First, it must reveal the mechanisms by which markets, and their stakeholders, are affected. Second, the framework must have an empirical analogue that establishes how the impacts can be quantified from observed market interactions.

A framework designed around a carefully constructed economic model can provide the means to address both considerations. Through theoretical constructs that lend themselves to graphic portrayals of market outcomes, one is able to characterize how the interaction of suppliers and consumers determines the explicit or implicit price of capacity. This is a critical step toward establishing the value of demand response when used as a source of capacity compared to its use as an offset to capacity.

Such a model allows one to trace through systematically how demand response the changes capacity prices, and identify who gains and who loses. Methods are available for estimating demand and supply relationships, and the corresponding equilibrium prices of capacity. In turn, it is possible to construct models to simulate the monetary outcomes of alternative forms of demand response under a variety of electricity market designs and institutional arrangements. These lead to empirical analogs that can be constructed to measure the performance of demand response programs that are implemented in electricity markets.

The discussion that follows reviews and synthesizes the body of research that addresses how to ascertain the value of demand response and establishes the foundations for a robust valuation framework that is developed in subsequent sections.

Unique Characteristics of Electricity Markets Affecting the Value of Demand Response Resources

The value of the resources committed to these various demand response programs is also inextricably tied to the unique nature of electricity markets. To quote Borenstein (2005, p.317), "[E]lectricity is not economically storable, and production is subject to rigid short-term capacity constraints. Because demand is highly variable, this means there will be times when there is plenty of capacity, and the only incremental costs of producing electricity will be fuel and some operating and maintenance (O&M) costs. At other times, the capacity constraint will be binding [there are no more units available for dispatch], causing the incremental cost to increase greatly and wholesale market prices to rise."

The situation is further complicated by the joint nature of production in electricity markets.[3] That is, in addition to supplying the energy commodity, the market must also ensure the security of electricity supply, which, according to Oren (2005), has been one of the overriding concerns in the restructuring of electricity markets. The National Electric Reliability Council (NERC) has defined reliability as the degree to which the performance of the components of the electrical system leads to power being supplied to consumers within acceptable standards and in the amount desired.

Since NERC's definition of reliability embodies the notion of "obligation to serve," Oren (2005) goes on to argue that NERC's definition may be out of step with deregulation and the creation of competitive markets. It is not surprising that the mix of system attributes needed to ensure the reliability of electricity supply has different technical and economic implications under alternative market structures. This may be particularly true if one is now in an environment where the notion of "obligation to serve" is replaced by "obligation to serve at a price". We return to this distinction below.

As discussed above, the concept of reliability is tied both to system adequacy in the long term, and to the short-term issues related to system security. As both Oren (2005) and Boisvert and Neenan (2003) observe, system security must be further distinguished because of its "public-good" aspects.[4] It is not possible to exclude those customers unwilling to pay for spinning

---

[3] In the economic literature, "jointness in production" of two or more products implies that they are economically interdependent (Beattie, *et. al.* 2009). For example, two products are economically interdependent if a change in the price of one product (i.e. p_j) affects the quantity supplied of the other (i.e. Y_i), and there are three distinct and instructive cases. If Y_i^* is the optimum (e.g. profit maximizing or cost minimizing supply of output Y_i), then:

- If $\partial Y_i^* / \partial p_j < 0$ (the demand for good i goes down the the price of good j increases), then Y_i and Y_j *are economically competing*;
- If $\partial Y_i^* / \partial p_j = 0$, (demand for good i is unaffected by the price of j) then Y_i and Y_j are economically independent; and
- If $\partial Y_i^* / \partial p_j > 0$b(demand for good i goes up when the price of j increases), then Y_i and Y_j are economically complementary.

There are three major causes of joint production. The classic definition of joint products refers to things that cannot be produced separately, but are joined by a common origin or non-allocable input, and are produced in fixed, or nearly fixed proportions. The second cause is a technical interdependency implying that production of either output depends not only on the amount of the factor allocated to this output, but also on the level of the other output. The third cause is due to an allocable fixed factor. In this case, the total amount of a factor used in the production of each output can be identified, but the total amount of that factor available to the enterprise is fixed in the short run. Spulbur (1989, pp. 114-115) discusses the conditions that minimum-cost cost functions must satisfy in order for production to be non-joint. Let there be i = 1,…, m products, Y_i, then a firm's technology is non-joint if and only if the cost function can be written as the sum of stand-alone costs, that is, $C(Y,w) = \sum_{i=1}^{n} C(0,...,Y_i,0,...,0;w).$,

where Y = (Y_1,…,Y_m), and $_w$ is a vector of input prices. An important result is that the technology is non-joint, there are no economies of scope, and production may be organized efficiently with single-product firms. For multiproduct production to yield cost efficiencies, there must be returns to common or joint production of outputs. As a simple example, suppose there is a cost function for two good in which there is a fixed cost of an input used by both production processes. Then, if a regulator is to set prices for both products, there is the issue of how to allocate the fixed cost between the two products.

[4] Public goods are defined as those that exhibit both consumption *indivisibilities* and *non-excludability* (Tietenberg 1998). *Non-excludability* refers to the circumstance where, once a good is provided, even those who do not pay for or provide it in some other way, cannot be excluded from enjoying the benefits it conveys. Consumption is *indivisible* if one person's consumption or availability of the good does not diminish the amount available for others;

reserves, and other resources from the level of system security they provide. Thus, customers and Load Serving Entities (LSEs), acting only in their own self- interests, would not purchase sufficient reserves to maintain system security at an acceptable level. It is precisely for this reason that ISO/RTOs that operate competitive markets for electricity are charged with the responsibility of securing the electrical system.

A similar argument applies to vertically integrated and locally regulated electric utilities. Because of the unique nature of demand and supply in power markets and the adverse consequences of market failure caused by the public-good nature of the system-wide reserves, Boisvert and Neenan (2003), Stoft (2002), Oren (2005), and others argue that appropriate levels of system security may not be properly met through market forces along. They must be determined administratively, centrally managed and funded through mandatory charges.

Once the level of reliability has been determined and the corresponding resource adequacy requirement established, the way in which that requirement is met can and does differ considerably. These differences, in turn, may affect the value of demand response that functions in some fashion as capacity. The primary source of the differences is the structure of the electricity market. Hence, the discussion that follows focuses on specific market designs and examines how a particular market design affects the value of demand response.

Economic Efficiency in Wholesale Electricity Markets

Demand response programs are designed to provide participants with price or other incentives to change electricity usage. Therefore, an examination of the effects of demand response programs on electricity markets begins initially with a comparison with the economic efficiency (or lack thereof) in electricity markets conventionally operated, wherein a monopoly utility provides electric service to all consumers at prices that reflect the average cost of supply. Under these arrangements, most consumers face fixed, uniform prices despite the substantial hour-by-hour variation in wholesale electricity prices due to the constantly changing supply-demand equilibrium. This discordance results in the less than optimal use of societal resources, and provides a starting point of comparison against which to measure the effectiveness of demand response as a remedy to resource misallocation.

The evaluation of demand response usually begins by characterizing the performance of an alternative market for the electricity commodity, wherein consumers pay prices that reflect the prevailing marginal cost of supply. In so doing, the resulting market equilibrium is consistent with the first-best, social welfare maximizing outcome.[5] For simplicity, and without loss of generality, most theoretical analyses are based two levels of demand—peak and off-peak.

---

thus, there is a potential problem with free riders. Since no private provider of a public good is able to capture all the benefits from its provision, the reliance on private markets to provide them will lead to levels of production that are below the socially optimal level. Some type of public intervention is required to ensure that production reaches the socially optimal level. Field (2001) points to some well known examples: lighthouses and radio signals have substantial "public good" aspects, as do national defense and clean air.

[5] Boisvert and Neenan (2003), Borenstein (2005), and others have provided recent analyses of this kind, but the inefficiencies or social deadweight losses due to distortions in what otherwise would be competitive markets (e.g., price floors, ceilings, and supports, taxes, subsidies, quotas, etc.) are well known (e.g. Just, *et al*. 2004). Those social deadweight losses due to fixed prices in electricity markets have been well understood for some time as well.

Where customers face fixed retail prices, it is easy to depict the social welfare losses during periods of peak demand. Customers use electricity up to the point where the value received from the use of the last unit is equal to its administratively determined fixed price. This value is well below the marginal cost of serving this last unit of load during periods of peak demand. It is only when customers are charged a price equal to the marginal cost of serving load that the competitive equilibrium is reached, thus eliminating any deadweight (resource) loss. Furthermore, there are social welfare deadweight losses during off-peak times as well.

In off-peak times, the price customers pay is above the marginal cost of supply from a social point of view, customers should be charged a lower price so that they would expand consumption up to the point where the value of an additional unit falls to the marginal cost of generation.[6] Conversely, a high peak price is warranted to abate demand to the level that achieves socially optimal resource allocation, both in the electric sector and economy-wide.

The size of the deadweight losses that can be eliminated through time-differentiated pricing depends critically on the price elasticities of the supply for electricity relative to the price elasticities of demand. This is certainly one reason for the recent heightened interest in the empirical estimation of price elasticities of demand and demand response, although initial efforts to estimate these relationships date back at least to the late 1970s and early 1880s.[7]

Using this paradigm, Borenstein (2005b) treats real-time pricing (RTP) as a "gold standard", against which all demand response programs should be judged. RTP involves sending a new price to consumers every hour to reflect very closely the prevailing cost of supply. RTP achieves greater efficiency than simpler peak/off-peak pricing regimes because, although time-varying prices can send appropriate signals, the situation is made complex by the uncertainty about supply and demand in advance of any given period. Thus, the efficiency of a program of time-varying prices depends directly on its granularity, the frequency with which prices are changed, and is inversely related to it timeliness, the time lag between when a price is set, and when it becomes effective.

A system of fixed retail prices is clearly at one extreme, with low correspondence at times between price and current supply cost, while RTP is at the other extreme, contemporaneous correspondence.[8] While much of the debate over RTP is on granularity, Borenstein (2005b, p. 325) also argues that timeliness is equally important—even to the point of whether RTP prices are set an hour or a day in advance. This is due in large measure to the fact that retail price programs offer prices as "requirements contracts"—the customer can choose to buy only what is desired at the time of delivery. In stark contrast, contracts for advance purchases at wholesale

---

However, with the increased interest in demand response programs, these kinds of analyses have been reformulated to act as a starting point from which to assess the effects of a broader range of DR programs.

[6] The essential components of the analyses for both peak and off-peak periods by Boisvert and Neenan (2003) are summarized in Appendix A of this report.

[7] Neenan and Eon (2008) provide an excellent discussion of the various concepts related to quantifying how electricity customers respond to changes in the price of electricity in the face of other compounding factors. They include in the reference list a number of older and more recent efforts to measure these responses empirically.

[8] RTP can be constructed by setting a new price each hour, but system dispatch operations allow for setting prices more frequently, even every five minutes or so. However, the marginal gain in price/cost correspondence has to be weighted against the added cost to consumers in being able to adjust usage on such short notice. Most proponents of RTP-type pricing argue that hourly prices strike a good balance between charging prices that reflect changes in marginal supply costs and the customers' "transactions costs" of adjusting usage on short notice.

specify both the price and quantity; deviations between customer consumption and advanced wholesale purchases are settled at spot-market prices.

Measured against the "gold standard" of RTP, Borenstein (2005b) argues that the increases in efficiency from other forms of time varying prices pale in comparison. His arguments are primarily conceptual, with no reference in that paper to attempts at empirical quantification.[9] Although used to some extent in the United States for large industrial and commercial customers, Borenstein (2005b) argues, for example, that Time-of-Use (TOU) rates lack the timeliness to capture any short-term variation in supply-demand balance (See Barbose, *et al*. 2004 and 2006). Because of the lack of granularity, the TOU prices also fail to reflect long-term expected variation in wholesale market prices. Due to the blunt, dramatic price changes, Borenstein argues that interruptible demand programs (where customers are paid to reduce load when called on to do so) are nearly at the opposite end of the spectrum relative to RTP; their major benefit is that they offer customers some insurance because they can be called to reduce load only a pre-specified number of times.[10]

Critical Peak Pricing (CPP) programs, which allow utilities to call a limited number of high priced hours, Borenstein argues, have advantages over TOU prices even with customer charges, and also have some of the advantages of RTP because retail prices can be varied with the wholesale market prices. Finally, Borenstein (2005b) argues that real-time demand reduction programs (DRP), which attempt to deal with the idiosyncratic hourly price variation of systems in stress and give customers incentives to respond, suffer because there is no reliable baseline against which to pay for performance, and the money has to come from somewhere—perhaps through higher general rates to meet revenue requirements. Thus, in his view, they appear as imperfect substitutes for CPP, but require almost the same level of metering technology.

Borenstein's perspective, which is echoed by others, is that the ideal form of demand response involves bringing forth the forces of supply and demand in the same ways that other sectors work. Suppliers offer prices for energy (kWh) and consumers decide at what level to consume based on those prices. The resulting energy prices will clear the market, and these prices serve to direct investment in capacity to serve load. In this regard, rising prices are a signal of potential shortages, and if they persist, they should induce investments in new capacity. Once new capacity is added, prices are restored to equilibrium levels that equate the marginal value of electricity in consumption to a lower marginal RTP price. This economic model recognizes and embodies explicitly the joint nature of the provision of energy and system adequacy and security. Price serves to allocate available capacity and to direct investments in generation to achieve an acceptable level of adequacy.

---

[9] In related research, Borenstein (2005a), however, does attempt to simulate the long-run equilibrium impact of demand response (defined generally as RTP) in a stylized world in which generation capacity adjusts immediately to reduced electricity consumption due to demand response. Using simple simulations based on what he characterizes as realistic parameters, he demonstrates that the magnitude of efficiency gains from RTP is likely to be significant even if demand is very inelastic, and that "time-of-use" pricing is likely to capture a very small share of the potential efficiency gains.

[10] Boisvert and Neenan (2003, pp. 15-17) demonstrate that the deadweight loss that can be avoided in the small number of times customers are asked to reduce load can be substantial. Thus, as long as this payment is smaller than the social deadweight loss that is avoided, social welfare is unequivocally improved. The size of the welfare gain is an empirical question, and it depends on the relative sizes of the price elasticities of supply and demand for electricity during these events. The situations in which both the supply and demand for electricity are very price inelastic are those in which the welfare gains are likely to be positive.

Unfortunately, markets that universally employ RTP seem to be only an aspiration. Most electricity markets have effectively bifurcated the provision of capacity and the setting of prices for energy usage. The result is that there are separate markets and prices for capacity and energy in which demand response can play a role, but not necessarily an equally effective one.

The next section describes how capacity adequacy is defined and how the obligation is met in U.S. electricity markets. This discussion is the essential backdrop needed to develop a framework to establish the value of demand response as a capacity resource in the sections that follow.

4
PROVISION OF RESOURCE ADEQUACY IN U.S. ELECTRICITY MARKETS

Electricity is generated on demand because there is very little storage capacity available. For this reason, electricity markets, in addition to supplying the energy commodity itself, must also ensure the reliability of electricity supply. As discussed above briefly, the concept of reliability is tied both to system adequacy in the long term, and to the short-term issues related to system security. Our purpose in this report is to analyze how demand response can be used to contribute to system adequacy. To do that effectively, however, we must first draw sufficient distinction between *capacity adequacy resources* and *operational reliability (or security) resources* to demonstrate that the way in which demand response can be used to contribute to system capacity differs from the way in which it can contribute to system reliability. Having made this distinction, we proceed to develop a framework of ascertaining how demand response influences capacity supply costs

Electricity Supply Assurance

Because electricity is essentially not storable, electric systems are designed to provide reliable service by taking into account the consequences of the unavailability of a generation unit, along with the loss of a key transmission asset or a surge in demand. In some instances, the latter two outcomes can be of greater consequence than the loss of a large generation unit. This is especially true for the loss of a key transmission line; if the power can not be delivered, then the availability of the unit is not the cause of concern.

This reliability assurance can be thought of as having three interrelated components: 1) establishing a measure of reliability, 2) defining what level is adequate to meet consumers' electric service demands, and 3) specifying how that goal is accomplished through operational protocols. The provision of reliable service is accomplished in two stages. The first involves anticipating (or forecasting) demand and then building the generation and transmission assets and systems to meet those requirements. The second involves using the generation available effectively to meet electricity needs on demand. In the first stage, investment decisions are made that involve large capital expenditures on generation facilities that can take years to build and may be operational for decades. This long-term perspective is referred to as *capacity adequacy*. The second stage is concerned with a short-term perspective and the requirement that demand can be met instantaneously and continuously. This task is referred to as operational *reliability or security.*

Capacity Adequacy

The reliability of an electric system is defined in terms of how often consumer demand for electricity is expected not to be met, typically stated in terms of the percentage of time that not all load is served over a specified planning period. It is posed in terms of an expectation because this metric is used to determine how much generation and transmission capacity to build to meet

forecasted demand. The more capacity that is available for dispatch, the lower is the likelihood of circumstances where generation is insufficient to meet demand.

One conventional measure of reliability is the number of days in ten years that electric demand is not fully fulfilled. One day in ten years is the industry standard. Put differently, the system should be designed such that on no more than one day over a ten-year period will demand for electricity not be met. This industry standard translates into a service interruption with the likelihood of three-tenths of one percent.

More generation reduces the likelihood of a shortage and service interruption, but adds to the costs that consumers must bear. Thus, a reliability standard should reflect consumers' willingness to pay for capacity. Using that criterion, the level of reliability would be defined by the generation capacity in an amount that equates its marginal cost to the marginal value consumers realize from the corresponding additional unit of reliability. At this point, the addition of another increment of capacity would not be justified by the consumers' willingness to pay. Having less capacity (and therefore lower reliability) would leave consumer demand for capacity unfulfilled.

The source of that one-day-in-ten standard is not clear. There is no indication that it was developed from any systematic comparison of the marginal cost and value of additional generation in terms of the change in value to consumers of reliability that is likely to result. The procedures for establishing the marginal cost of an addition to the generation portfolio that serves an electric system (or market) are quite well known. However, the procedures to establish the marginal value to customers of additional reliability are considerably more complex and involve a large degree of subjectivity.

Complexity arises from the diversity of how electricity is used by heterogeneous consumers and therefore how it is valued. The loss of power to the air conditioner of a household or office building, for example, is an inconvenience to the occupants. However, that value of that loss pales in comparison, most would agree, to the loss of power to critical infrastructure such as hospitals and health care facilities. Despite the difference in the severity of the consequences in these two rather extreme examples, establishing a monetary value for these losses in either case is a challenge, as it involves highly subjective assessments.

Business losses that result from the curtailment of electric service may be easier to monetize because they consist of such things as lost sales or the destruction of products or perishable inputs of production. But, even in these cases, the levels of losses suffer differ considerably depending on the nature of the business, the time when the outage occurs, and for how long it lasts. Considerable research has been undertaken to quantify the value of electric service outages. Nevertheless, there is still no agreement on acceptable valuation methods that can be applied universally to test whether the capacity adequacy standard is indeed optimal.[11]

The capacity adequacy standard enjoys near universal acceptance, despite its provenance. It is generally treated as a fixed obligation. Electric system planners determine what capacity (generation and transmission) is needed over the planning cycle (10 or more years) given their estimates of load growth (level and profile), the capabilities of the existing capacity portfolio, unit retirements, and other factor that influence electricity demand and supply.

---

[11] The results from a meta-study of the estimates of the value of lost load are contained in Sullivan, *et al*. (2009).

The adequacy requirement by convention is determined by the level of peak demand, defined by the highest forecasted hourly system load. In some cases it is a function (for example the average) of several of the highest loads on several days during the year, such as those associated with the months of the peak electricity demand season. Historically, the determination of this adequacy requirement does not take into consideration the extent to which electricity consumers would be willing to purchase greater reliability, or to accept a lower level of reliability. There are, however, some important exceptions to this convention, and these are discussed below.

A variety of capacity planning models are used to ascertain how much capacity, and corresponding transmission assets, are required to achieve the (no more than) one-day-in-ten outage expectation. These models are used to simulate how available system resources would be dispatched to meet forecasted demand over a range of system conditions. The models produce a measure of adequacy that serves as a proxy for determining the extent to which available capacity can meet the adequacy standard. When load grows, or capacity is retired, so that the adequacy standard can no longer be met, then the least-cost addition to capacity that is required to redress the shortfall is added to the portfolio. Through this process, the additions to capacity that are required are determined, and they constitute the capacity expansion plan. How that plan is put in operation depends on how the electricity market is organized and operated. Subsequent sections of this report address the extent to which such distinctions influence the level of the capacity requirement and the role that demand response can play in achieving it.

The end result from the application of this adequacy standard is that capacity in excess of the expected peak load is provided. To meet the adequacy requirement, the system must have generation in excess of the peak load in order to continue to serve load when one or more units become unavailable. This excess capacity is called the reserve margin. The reserve margin requirement is usually between 12% and 18% of expected peak load. Differences within this range reflect the differential character of a utility or market's portfolio of generation (the size and scope of the largest units, and perhaps the fuel they use to generate electricity), the nature of the demand to be served (for example, the extent to which peak demand is weather sensitive or growing), and the determinations by regulators and stakeholders regarding what constitutes sufficient capacity to achieve an acceptable level of expected service reliability.[12]

System Operational Reliability or Security

The capacity adequacy requirement establishes the stock of generation (and transmission) capacity that must be available to serve system electricity load during the year based on the peak demand forecast and the operating characteristics of the units. That load must be served on demand, instantaneously and continuously, which involves making provision for circumstances whereby things do not go as planned. These contingencies include such things as the loss of a generation unit or units of such a magnitude that the consequences reach far beyond the inability to serve load equal to the capacity of the lost units. The nature of such contingencies leads to instability of the entire electric system; as a result, other units could go off line and in an extreme scenario, parts of the system or entire system could fail leading in turn to wide spread customer blackouts. The maintenance operational reliability or security is achieved through the provision and operation of sufficient resources to address and mitigate these potentially catastrophic circumstances.

---

[12] Investments in energy efficiency are sometimes posed as alternatives to building capacity to serve load.

In this sense operational security is assured through the establishment of operating reserve margins that define the amount and nature of generation units that are on stand-by, ready to provide energy to the system (or serve other roles such as voltage maintenance) in the event that there arises a set of circumstances that could have drastic consequence for the delivery of electricity to consumers.[13]

The operating reserve margin is determined by the system contingency that is the most detrimental to short-term security. Such contingencies can be reflected in the loss of the availability of the largest generation unit, or a combination of units and transmission delivery capability. The system operating requirement embodies the idea that the system must be designed and built so that it can withstand the largest single contingent event that would threaten its ability to serve demand. If this is so, then the system operating requirement is met.[14]

Capacity requirements to maintain the operational system reliability are defined according to how fast the generation would be required to come on line. Some contingent resources must be able to alter their output quickly (within a few seconds) to respond to changes in load that occur faster than the units on line to serve demand can adjust their output. Others are truly contingent; they are stand-by units purposely not providing energy to serve exigent electricity demand. Instead, they are held in reserve so that they are available to come on line when a contingency arises.

There are other forms of contingent response such as the curtailment of service to customers. In some instances, it may well be preferable or necessary to curtail service to some customers in order to avert the possibility of a wide-spread black out for a significant number or all customers.

The technical details of how this operational system security is provided and managed are covered elsewhere. Our primary purpose here is to draw a sufficient distinction between capacity adequacy resources and operational security resources to demonstrate that the way in which demand response used to contribute to system capacity must differ from the way in which it can contribute to system reliability. The discussion that follows focuses on demand response as a capacity adequacy resource. However, before engaging in that discussion, it is instructive to consider the implications of a system wherein capacity adequacy or reliability is not considered as a fixed requirement, but rather is determined by consumers' willingness to pay the cost of its provision.

The Demand for Adequacy and Reliability

The convention in the electricity industry has been to treat the demand for capacity adequacy and operational reliability as completely price inelastic. Consumer demand for these two attributes of the electric system are characterized as being invariant to the cost of providing the specified levels of generation capacity, and associated transmission delivery assets, to meet those requirements. The implications of this portrayal of consumer demand for capacity and reliability have received increased attention recently in part due to the formation and operation of wholesale electricity markets.

---

[13] The immediate consequences are the damages to generators that can result from a collapse of system stability. This leaves these generators unavailable to provide power to consumers; the result is even grater damages.
[14] For details, see Kirby, *et al*. (2002) and Kirby and Hirst (2000).

Some wholesale markets for electricity specify the requirements for capacity adequacy and operational reserves and operate markets to bring buyers and sellers together to help them meet that obligation. When these markets began to determine the prices for capacity, which heretofore had been established by administrative hearings, and those prices were posted publicly, the escalating costs of providing for these services became a serious source of contention.

Almost from the outset, market prices for capacity were highly volatile in the wholesale capacity markets in New York and the Mid-Atlantic states served by PJM. This volatility, as depicted in Figure 4-1, was in large part due to a fundamental characteristic of capacity supply. Along these supply curves, prices rise modestly as the quantity supplied increases up to the last few last few increments of available supply. Then, the supply price rises dramatically in response to relative small changes the amount supplied. The shape of this supply curve is often likened to that of a hockey stick, where the handle represents the last and steepest segment of supply.

The sharp rise in capacity supply costs reflects the higher ownership cost associated with marginal supply units, such as peaking units, that are rarely used, and therefore generate less income for being dispatched than do base units that operate almost continuously.

In addition, market operators and stakeholders were also increasingly concerned that capacity prices, for both adequacy and reliability, were too low to attract new investment. If this were indeed true, the result would be a shortfall of capacity that would ultimately lead to much higher capacity costs to consumers and could potentially compromise adequacy and reliability.

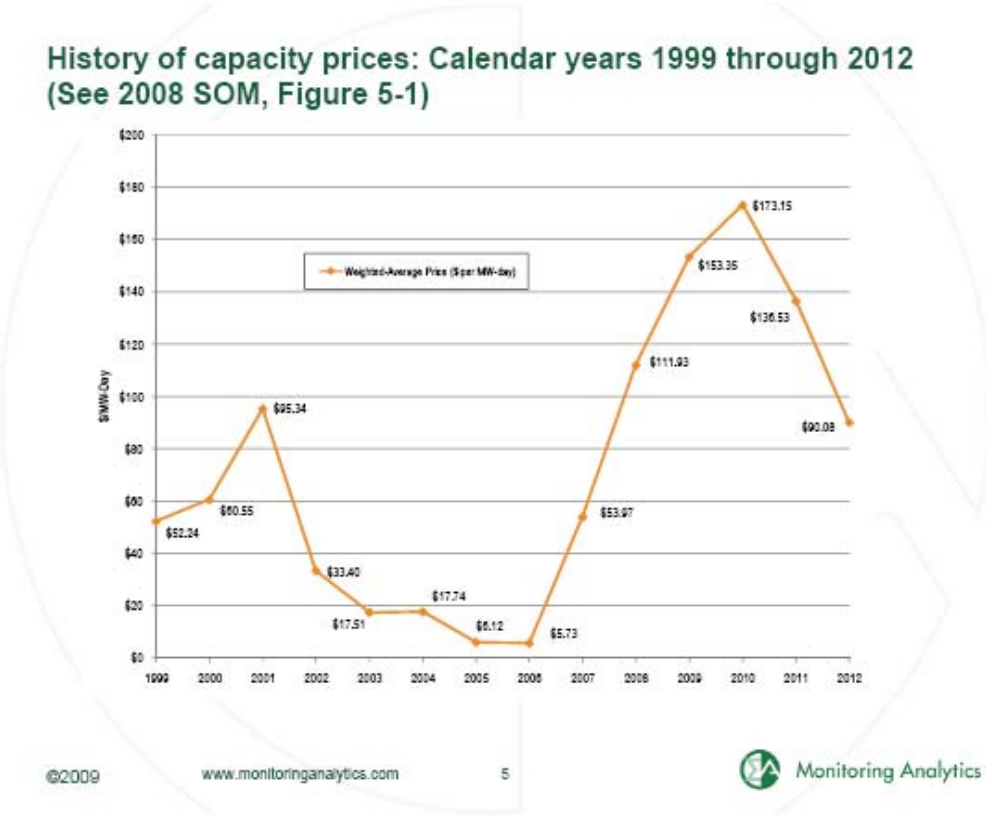


Figure 4-1

NYISO and ISO-NE, addressed these concerns by characterizing the demand for capacity and/or reliability as a function of the price of supply. Instead of a vertical demand curve, they represent capacity demand as a downward sloping curve over a portion of the range of supply availability. Figure 4-2 illustrates this concept.[15]

The demand curve for capacity is vertical (not price responsive) over the initial range or segment of the price/quantity space, indicative of the essential role of capacity for many consumers. Then, at some price it can be horizontal over some range, indicating that the market will pay a constant price for any amount of capacity within this range. Finally, the demand curve becomes downward sloping at some specific MW level (or percentage of required capacity). At this point, the amount of capacity the market is willing to buy increases as the price falls.[16] Under this market structure, the intersection of supply and demand for capacity still determines the market price, but price volatility is reduced somewhat because there is a relatively small impact on capacity costs for small changes in supply at all but the very highest level of supply. As a result, consumers are provided greater reliability when the equilibrium in the market occurs within this downward sloping portion of the demand curve.

Proponents of this administratively-set downward sloping demand curve for capacity tout the fact that this characterization of the demand for reliability as price responsive is consistent with studies that demonstrate that many consumers would pay more for increased reliability that the conventional standard implies. Critics, on the other hand, argue that the demand for capacity curve is derived administratively and not based on revealed or stated consumer preferences. They argue that these demand curves may not correspond to any actual consumer willingness to pay for increased reliability. The upper portion of the demand curve is set so that the price of capacity is approximately equal what has been determined to be sufficient to attract new generation capacity. The downward sloping portion is more of an intellectual construct than it is a realistic characterization of consumer demand for capacity. These are hypothetical constructs that are more reflective of assumptions about what the cost or capacity needs to be to attract new supplies than an expression of consumer's willingness to pay for them.

Some centralized electricity markets have acknowledged the fact that not all consumers have been satisfied with the commonly imposed and conventional level of capacity adequacy (and consequent level of reliability). Some have then gone on to make provisions for consumers to become de factor suppliers of their own capacity. This is accomplished by allowing customers to

---

[15] For a detailed description, see Stoft (2002) and Hobbs, Inon, and Stoft (2001).

[16] In a 2007 report, NERA provided estimates for the parameter of the ICAP demand curve for the NYISO. Separate demand curves are formulated for New York City, Long Island, and the rest of New York State. These recommended demand curves were compared to the ones in force at the time, and these differed substantially for these three areas. For upstate New York, for example, the demand curve in force at the time of the study is vertical at 80% of required capacity down to a price of about $148/kW year. At this price it is horizontal for capacity between 80% and about 90% of required capacity (including the reserve margin). From that point on, it is falls linearly and intersects the price axis at a zero price at about 112% of required capacity. At 100% of required capacity, the price is $74/kW year. In New York City, the demand curve is vertical at 80% of required capacity down to a price of about $280/kW year. It is horizontal between 80% and 85% of required capacity (including the reserve margin) at a price of about $280/kW year. From that point, the demand curve falls linearly and hits a price of zero at 118% of required capacity. There is a price of about $139 at 100% of required capacity. On Long Island, demand is vertical at 80% of required capacity down to a price of about $250/kW year, and there is no horizontal portion to the demand curve. Demand then falls linearly everywhere between 80% of required capacity to 118% of required capacity. This yields a price of about $84/kW year at 100% of required capacity.

become de factor suppliers of their own capacity. This is accomplished by allowing customers to offer load curtailment capabilities, which are referred to as demand response, as a supply resource for capacity.

In these markets, consumers can participate in the capacity provision process. They do so by selling the demand response capability to load serving entities and electricity retailers responsible for meeting capacity adequacy requirements. The purchase of this demand response can serve as an alternative to the purchase of capacity from a generator. In some cases, consumers can also offer their indigenous demand response capacity to a centralized capacity auction operated by the ISO/RTO market operator.[17] It both cases, if the consumer's capacity offer is accepted, the customer is paid the market price (or close to it) for capacity, in a way similar to the way in which generators are paid for capacity. In return, the customer has sacrificed some its reliability assurance.

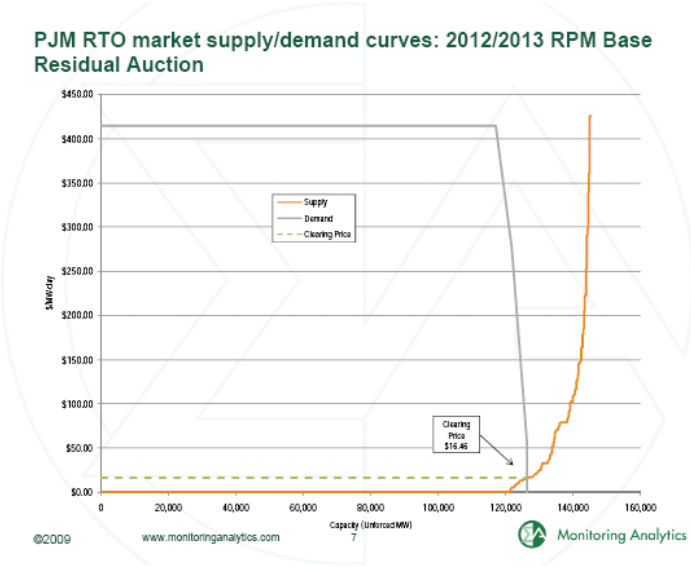


Figure 4-2

Since the early 1970s, some vertically integrated utilities have used interruptible service tariffs to offer consumers discounts to accept a lower level of service reliability. Load pledged to this service reduces the system capacity requirement almost megawatt for megawatt. In other cases, consumers that reduced their peak usage on the peak day realized reduced capacity costs the subsequent year; this amounts to receiving a capacity payment equal to the demand charge they paid. Some have proposed that all consumers should have to nominate each year how much firm power they want to purchase, and that would be amount of capacity that would be acquired. However, this scheme has yet to be fully implemented, in part because its efficacy required determining when to curtail the usage of those that selected a lower level of reliability.

[17] For details on how these markets operate see: ISO/RTO Council Markets Committee (2007a and 2007b) and Heffner (2009).

Demand Response as a Capacity Adequacy Provider

Electricity markets are being restructured to acknowledge explicitly that not all consumers require the same level of service reliability. Based on this recognition, some centralized wholesale markets allow consumers to offer demand response as a capacity supply alternative to generation capacity, and in return, they receive the market clearing price. Vertically integrated utilities and cooperative and municipal utilities with similar market responsibilities offer customers reduced rates or payments if they agree to a lower service standard.

These initiatives share a common goal: equating the level of reliability provided to what consumers are willing to pay for reliability. As these markets gain experience with these types of program alternatives, one would expect that the payments made for demand response would gravitate to a common value, or band of predictable values. At any point in time, differences may arise in what demand response is worth across markets. The variations in the current supply and demand conditions across these markets would rationalize these differences.

Today, the payments offered for demand response vary considerably among markets, and they exhibit large year-to-year changes in the same market. For example, consumers in New England that provide capacity adequacy to the centralized ISO-NE market (because they were selected though an auction process) in the form of demand response receive about $30/ kW year. Most customers providing demand response to PJM in 2010 will receive $30/kW year, but some will receive over $80/kW year. Several utilities operate programs that pay $20-40/ kW year for participation in load control (AC, water heating, pool pumps) programs, while others offer no capacity programs, or limit participation in existing programs because the utilities claim that the value is zero.[18]

Ambiguity about the value of demand response acts as deterrent to the clear resolution of what the ultimate role might be for demand response as a capacity resource. Utilities find it difficult to design long-term programs in the face of uncertainty, and because of the prospect of being second guessed if they take a longer-term position on demand response. Unstable or unpredictable capacity payments result in customers being reluctant to make investments in technology that would enable them to participate in demand response capacity programs, or participate at a high level. Vendors of new enabling technology and curtailment service providers are equally stymied as they lack sufficient information on which to make affirmative business decisions.

For these several reasons, a robust framework to assess the value of demand response as a capacity resource would contribute toward clarifying for all stakeholders the fundamental relationship between consumer demand for reliability and the cost of supplying those demands. It is to this task that the remainder of this report is focused. The framework relies on fundamental market principles derived from economic theory. Within this framework, one can explain how changes in supply and demand influence the market price for capacity, contribute to market efficiency and square with the maximization of social welfare. The approach provides insight into how centralized markets value demand response as a capacity resource, and because the defining character of demand is universally applicable, the findings can also be extended to other market structures.

---

[18] These are examples of current payments. No comprehensive database of capacity payments for demand response is available.

5
A FRAMEWORK FOR VALUING DEMAND RESPONSE AS A CAPACITY RESOURCE

The preceding discussion has suggested why the market structure may be a primary determinant of the value of demand response as a capacity resource. The development of a framework to value demand response resources therefore must address these differences. One can begin by describing an idealized construct whereby each consumer is responsible for procuring the level of resource adequacy it deems sufficient to meet its needs. In turn, this demand for capacity and the corresponding supply of capacity determine the prices consumers pay for capacity. This is an idealized market in that it extends to the partitioned capacity market the features of an all-energy market which achieves the highest degree of economic efficiency. One can place a value on capacity in this idealized market and then determine how the value of capacity differs from this norm within other electricity markets that are in place today. This comparative analysis will make transparent any differences in the value of capacity particularly between ISO/RTO-run markets and markets run by vertically integrated utilities.

A Simple Market Model for Capacity

The stylized market for capacity is depicted in Figure 5-1. The vertical axis measures the market-clearing price for capacity and the horizontal axis measures the quantity of capacity available from generators.

For this market, it is assumed that the demand for capacity is fixed at D_M, equal to system peak demand plus a reserve margin as prescribed by resource adequacy rules. For simplicity, and also to abstract initially from the time dimension inherent in system planning periods of from six months to a year, one can assume that this demand for capacity is for one hour only. Much of the complexity introduced by explicit consideration of securing capacity over a typical multi-year planning period can be addressed by examining situations in which this one hour reflects different conditions for the supply and demand for capacity due to seasonal considerations or due to the different capacity demand distributions of customers by customer class.[19]

In this initial model, demand for capacity is fixed. Therefore the demand curve (the curve that maps out the amount of capacity demanded at each price) is depicted by a vertical line intersecting the horizontal axis at $Q_M = Q_G$. The supply of capacity from the generators, S_G, is a non-decreasing step function, resembling a "bid" stack for generators bidding into an energy market. The first step reflects the fact that base load plants are willing to supply capacity at a low price because of their low operating costs. Plants with higher, but still moderate operating costs are in the second step, while peak load plants are able and willing to supply smaller increments of capacity at higher and higher prices. Alternatively, the upward sloping supply curve reflects the different annualized costs of generation that a utility faces to secure capacity.

[19] This formulation is consistent with a system that establishes its resource adequacy requirement based on the system peak hour. More typically, the average of the peaks in the summer months is used to establish the adequacy requirement.

The length of each step in S_G is the maximum amount of capacity that will be supplied at the corresponding price. Thus, the right end point on each horizontal portion of each step (the hip) indicates the maximum combined amount of capacity that generators are willing to supply at the corresponding price. However, at the point where all generation capacity is committed, no more can be supplied at any price, and the supply curve becomes vertical. In Figure 5-1, this occurs at a supply of Q_{GMAX}, and results in a capacity price of P_{VG}.[20]

Market Equilibrium where Supply Fulfills Demand

The equilibrium in this stylized capacity market occurs at the intersection of supply and demand, the point labeled E_G in Figure 5-1, where the supply of capacity is equal to the fixed demand for capacity, $Q_M = Q_G$. Each unit of that capacity is paid a price of P_G. In this case, the demand for capacity is below the limit on generation capacity availability, Q_{GMAX}. As depicted, this equilibrium represents the typical case in that the market clearing conditions are satisfied (e.g. supply equals demand at a finite price). Typically in a well-functioning market, there would be sufficient capacity to meet demand. Furthermore, if the demand for capacity is truly fixed (e.g. customers are neither willing nor able to alter the demand for electricity), then this represents the socially optimal solution as well. This is so because there is no under-served or over-served demand for capacity, and hence there are no deadweight losses to society. Deadweight losses represent squandered resources that result when resources are not optimally allocated.

There are, however, two potentially disruptive situations in which the market does not reach this equilibrium. These situations are depicted in Figure 5-2 and Figure 5-3.

Market Equilibrium with an Abrupt Shortfall in Supply

The first situation, illustrated in Figure 5-2, is where for this stylized representation (a single hour) one or more generation units are out of commission. In this case, the outage is assumed to have affected units that were willing to supply capacity at a price of P_{OUT} or less. Consequently, the capacity that can now be supplied at this price is reduced by the amount in brackets in Figure 5-2, labeled 'Outage', which corresponds to the output of those units. Thus, at prices equal to or above P_{OUT}, there is a leftward shift in the supply curve (from S_G to S'_G in Figure 5-2) to in effect fill the void created by the absent units. The inoperative part of the original supply curve is now depicted as the dashed yellow part of the original supply curve S_G (Figure 5-2).

Because of this outage, the effective supply curve is now S'_G in Figure 5-2; it becomes vertical at a lower level of capacity, as indicated by Q'_{GMAX}. This new point at which the supply curve becomes vertical is to the left of the fixed demand for capacity, D_M (which has not changed). Due to the units becoming unavailable, there is insufficient supply of capacity (at any price) to meet all the demand for capacity. This disequilibrium in the market is caused by the unforeseen generator outages.

---

[20] As is often depicted for the supply curve for energy, this supply curve for capacity also has the "hockey stick" shape because as one nears system generating capacity, it is only the most expensive units that can supply additional capacity. As is seen below, and for the purposes of placing a value of demand response for capacity, this is an important characteristic of the supply curve for capacity due to the finite limit on generation capacity at any point in time.

Market Equilibrium with a Abrupt Increase in Demand

The second situation, depicted in Figure 5-3, is where there is a shift to right in the demand for capacity—from D_M to D'_M. This might be the result of a spike in temperature that causes system load to rise abruptly. In this case, nothing has happened to supply. The supply curve is the same (S_G) as in Figure 5-1. This market disequilibrium is now due to a change in demand. The demand curve remains vertical, but it has shifted to the right, beyond the point of maximum capacity, Q_{GMAX}.

Although for quite different reasons, the demand for capacity in both cases now exceeds the maximum that could be supplied by generators at any price--Q'_{GMAX} and Q_{GMAX}, in Figures 5-2, and 5-3, respectively. Under both circumstances, there is no finite price at which supply can equal demand and the market will clear serving all of demand. In these cases, the market rules would have to be suspended and prices would somehow be set administratively.[21] Furthermore, in the absence of rationing rules that reflect consumers' willingness to pay, any load shedding undertaken to reestablish system-wide reserve contingency margins would have to be accomplished administratively.[22] Since there is no way to ensure that this load shedding will be assigned to those customers who place the lowest value on these last increments of load, there is real potential for allocative inefficiencies. The social deadweight losses could be substantial.

As is seen below, it is not surprising that it is in these extreme situations where the value of additional capacity supplied by demand response customers is highest, but that's getting a bit ahead of the story. To make this analysis transparent, we must first introduce demand response into a market that is functioning normally. For the analysis to be meaningful, we must presume that there are indeed some customers willing to supply capacity through participation in a demand response program of this type. Recent experience, particularly in markets run by ISO/RTOs, suggests that this is a reasonable presumption.

Supplementing Capacity through Demand Response

In this section, we introduce into the market for capacity a new source of supply from demand response program participants. Consumers can lower their capacity requirement by agreeing to reduce that amount of load when called on to do so. In this formulation, the supply of capacity by generators remains unchanged. But, demand is adjusted to reflect consumers' willingness to pay for capacity based on what they would have to pay for capacity.

In is convenient to discuss this new source of supply in two steps. The first is to contrast the supply curve for capacity and the demand response participants from the overall demand for capacity under these conditions. Then, we can overlay this new supply curve for capacity by demand response participants on the supply curve for capacity by generators. When these two

---

[21] During the Northeast blackout of 2004, for example, ISO/RTOs had to set prices for energy and ancillary serves administratively because the market pricing algorithms were not designed to handle system restart under conditions where supply is substantially below demand.

[22] Provisions for load curtailment are part of dispatch rules, but they seldom reflect bids by consumers to avoid an outage. California implemented programs to allow customers to specify, to some extent, where in the load curtailment order they would be placed. The NYISO requires that load acting as a capacity adequacy resource provide a strike price that determines the order they are dispatched during conditions when not all the available curtailments are needed.

supply curves are then overlaid on the demand curve for capacity, the effect of these additional capacity resources on the market becomes transparent.

For this exposition, we assume that consumers are required to nominate the amount of capacity they are willing to pay for (their election of firm power) and that the total capacity acquired is determined by comparing the collective willingness to pay with the cost of supplying capacity. The system operator allocates available capacity in each hour of the year according to the nominations, first allotting capacity to the highest valued loads and then on until all is allocated. In a well-functioning market, in most cases there is likely to be sufficient capacity to serve all loads. However, in some cases there will not be enough, and some consumers will be denied service; it will be those that chose not to purchase firm power that will be denied service.

This capacity rationing system has been referred to as a demand subscription service because it in effect allows consumers to subscribe to the level of firm service they want. In practice, its application requires that the system operator is able to curtail the loads of individuals selectively. This would be accomplished by means of protocols that relate system capacity availability (relative to demand) conditions directly to curtailment actions.

In Italy and Spain, for example, this is accomplished by installing in premises a demand limiting device so that the premise load can not exceed the prescribed level. In Italy, all residences are so limited to 3.5 KW. Spain allows occupants of the residence to select the level at which the load limiter is set (e.g. 3.5, 8, and 12 kW). Similar programs, refereed to as demand subscription programs, have been offered in some parts of the United States, but generally their availability has been limited, or the programs have been imposed to ration load in areas with limited delivery access. The most significant example of this capacity provision mechanism was implemented in Texas. Larger consumers received a discount from their utility in return for allowing the utility to install a device at their service entry. This device enabled the utility to shut off power completely and almost instantaneously. These consumers received a demand discount that was in effect equal to the capacity payment they otherwise would have made. A variation of this program is operated today by the system operator (ERCOT), but it is used to provide security services.

Supply Curve for Capacity from Demand Response Resources

For this stylized market for capacity, it is assumed that for this representative hour, the supply of capacity by participants in a demand response program is also an increasing step function, similar to that depicted for generators. This supply curve for demand response capacity resources is depicted in Figure 5-4, but it is done so in a rather unconventional way that needs a bit of explanation.

To understand the construction of the figure, it is critical to be reminded that for every unit of demand response capacity provided, there is a corresponding reduction in load.[23] Therefore, in theory at least, and for the right price, the maximum amount of capacity supplied by participants in a demand response program could be equal to Q_M, the fixed demand from Figure 5-4, *less* the reserve margin. In effect, because of the joint nature in production, the demand for capacity also defines the supply of capacity. This is seen most easily in Figure 5-4, by starting at D_M, and

[23] As mentioned above, this is a case where the two products, capacity, and load reduction, are produced jointly and in a one-to-one fixed proportion. Most demand response programs attribute more than an equal MW reduction in supply to reflect line losses.

measuring the supply of capacity from demand response participants by moving from right to left, as indicated by the arrow at the top of the figure. Moving in this direction, it is easy to trace out the supply curve. At a zero price, there is no supply of demand response capacity offered, and the demand for capacity is still the vertical line D_M, and therefore remains at Q_M. But at prices above zero, the supply curve contains three steps that reflect the provision of demand response. Each is described in turn.

To begin, if the price for capacity is P_{0DR} (Figure 5-4), then a maximum of Q_{0DR} units of capacity will be supplied into the capacity program. This initial step in this demand response supply curve for capacity could reflect a willingness to forgo discretionary electricity usage for air conditioning in an office building, or a manufacturing plant that would shift load to another time of day during periods when the curtailment is called for. Although employees may be inconvenienced, there may be little or no effect on sales or production. Thus, the price at which the capacity is offered into the market may be relatively low, but it may also be true that the amount of capacity that consumers are willing to forego would be rather modest as well.[24]

Similarly, at a price of P_{1DR}, a total of Q_{1DR} will be supplied. This will consist of the first Q_{0DR} units offered at a price of P_{0DR}, plus an additional amount of $Q_{1DR} - Q_{0DR}$ at the higher price of P_{1DR}. The third step on the demand response supply curve indicates that any additions to supply of capacity by demand response participants would come at a high price indeed (indicated by a break in the price axis). Put differently, this segment of the supply curve suggests that any additions to the demand response supply capacity (e.g. reductions in load between Q_{1DR} and Q_{DRMAX}) would only occur if demand response capability reached the very high price P_{VDR}.

Another way to conceptualize the outcome is to adjust the supply of capacity to reflect the reduced demand indicated by movements along the demand curve. Put differently, at each sequentially higher price for capacity, the indicated reduction in willingness to pay (demand for capacity) can be treated symmetrically as an addition to the available supply. From this perspective, we can also treat overall demand for capacity as being fixed (a vertical demand curve at Q_M) as was done originally. This is possible because now consumers are treated as suppliers of capacity though a demand response curve. They offer curtailments at various price levels which are the same demand reductions as were previously traced out. However, they are now counted as supply so the amounts are added to the supply curve at each price point, and demand remains the same; reflecting the demand for capacity in both the demand and supply curve would result in undersupply of capacity.[25] This new market perspective might be implemented through a requirement that all consumers submit bids for specified levels of capacity that they would be willing to supply at specific prices for capacity. These bids, in turn, would then be added to an already established supply curve for capacity by generators. It would also facilitate the interpretation of this bid curve as a derived demand curve for capacity.

To determine the market effects of this new supply of capacity by demand response participants, one need only superimpose the curves from Figure 5-4 onto Figure 5-1. This is done in Figure 5-5, where again, the supply of capacity by generators is read from left to right, while the supply of demand response capacity is read from right to left. However, because of the joint one-to-one

---

[24] There is ample evidence that at least some consumers are willing to forego consumption in return for an inducement that effectively reduces their capacity costs. For example, see: ISO/RTO Council, Markets Committee (2007a).

[25] ISO/RTO demand response programs follow this practice.

relationship between a consumer's supply of a unit of capacity and the simultaneous reduction in a unit of electricity usage, the consumers' demand for capacity can also be read from left to right on the demand response's supply curve for capacity.

As a consequence, the new equilibrium in this market for capacity is affected by the introduction of the demand response as a supply resource for capacity. This equilibrium occurs where the supply curve for capacity from generators crosses the demand curve for capacity by demand response participants. In Figure 5-5, this is at the point E^{14}_{G+DP}.[26] The equilibrium price is P^{14}_{G+DR}. Since the supply of capacity is now increased through demand response participation, it is not surprising that this new equilibrium price P^{14}_{G+DR} is below the price P_G which represents the equilibrium price in a market where only generators supply capacity. The capacity supplied by generators drops from $Q_M = Q_G$ without the demand response program to G^{14}_{DR} with the program. At this new equilibrium, participants in the demand response program now supply the remaining units of capacity, G^{14}_{DR} (the bracket labeled in Figure 5-5 labeled Q^{14}_{DR}). In doing so, they also simultaneously decrease the demand for capacity by an equal amount.

The Implications for Economic Efficiency

By conducting an analysis similar to that employed in comparing the economic efficiency of real-time pricing with flat rate tariffs in markets for electric energy, as discussed in Section 3, one can also examine the efficiency of capacity markets, with and without demand response resources. As stated above, if generators are the only suppliers of capacity (e.g. demand for electricity is fixed and does not respond to price changes so that no customers are willing to participate in a demand response capacity market), then the situation described in Figure 5-1 is a social optimum, and, absent any generator outage, or unexpected increase in demand, there are no deadweight social losses.

This is not the case, however, if there are customers willing to supply capacity through a demand response program, but they do not have the opportunity to do so. The social "cost" of there being no opportunity to participate in such a demand response program can also be seen by comparing the two market equilibriums depicted in Figure 5-5. As discussed above, point E_G is the equilibrium for a capacity market facing a fixed demand where capacity is supplied only by generators. In this market, the marginal cost of an additional unit of capacity is P_G. Based on the demand curve for capacity, the value to consumers of the last unit of capacity is P_{DR}, which is much lower than the supply cost. This indicates that customers are willing to supply some of the capacity through demand response program participation, and hence equilibrium E_G is inefficient.

It is only through the implementation of a mechanism whereby consumers can bid their willingness to buy capacity, or provide capacity in the form of curtailments to offset supply, that this inefficiency can be resolved. By allowing customers to supply capacity through demand reduction, the new equilibrium (e.g. the new intersection between supply and demand) in Figure 5-5 is at point E^{14}_{G+DP}. At this point the conditions for competitive market equilibrium are satisfied. The marginal cost of supplying a unit of capacity is equal to the marginal value in use (recall that the demand curve for capacity is observed by moving from the left to the right on the

---

[26] The superscripts and subscripts are designed to keep track of the different scenarios in the various figures. For example, the superscript 14 refers to the fact that this figure combines the generator supply curve from Figure 5-1 with the DR supply curve from Figure 5-4. The subscript G+DR refers to the fact that the market contains supply from both generators and DR participants.

demand response supply curve for generation). This occurs at a price of P^{14}_{G+DR}. Viewing this equilibrium from a different perspective, this new equilibrium is where the supply prices for capacity by both generators and demand response participants are equated, and both are equal to the value of capacity in use; this is as it should be to achieve allocative efficiency.[27]

Similar to the analysis of RTP programs, the measure of the deadweight loss avoided through the implementation of a demand response program for capacity is measured by the area under the supply curve for generators and the area under the demand curve for capacity between the two equilibrium quantities of capacity supplied by generators, Q^{14}_{G} and G_M. This is the area shaded in green in Figure 5-5. Using this framework, the deadweight losses are in monetary terms. Therefore, if one were to estimate supply and demand curves for capacity empirically, one could quantify the cost of not allowing consumers to express their demand for reliability.

It is clear from this analysis that a demand response capacity program can mitigate the social deadweight losses in normally functioning capacity markets with fixed demand where there is sufficient generation capacity to meet demand. However, the supply of demand response capacity can also reconcile the two situations, mentioned above in Figures 5-2 and 5-3, in which the demand for capacity exceeds the maximum that could be supplied by generators at any price. By making provision for curtailing customers with the lowest value for the last units of capacity, a demand response program may avoid deadweight losses that would otherwise occur by administratively (arbitrarily with respect to the value of reliability) implementing load shedding which may include the load of some customers with a high value for the last units of load.

A similar type of social welfare analysis for a demand response capacity program can be conducted for the two situations described above where there is a shortfall in capacity either because of a generator outage, or an unexpected outward shift in demand for capacity. One need only superimpose Figure 5-4 onto Figures 5-2 and 5-3, respectively.

This is accomplished in Figure 5-6 and Figure 5-7. In the situation in Figure 5-2, demand is above maximum capacity of generators due to some unforeseen outage. Thus, the market will *not* clear at any price. However, if customers are allowed to bid capacity into the market through a demand response program, the market will clear at a finite price of P^{24}_{G+DR} in Figure 5-6. Because the supply of capacity by generators is reduced at prices above P_{OUT}, this equilibrium price is above P^{14}_{G+DR} from Figure 5-5. The amount of capacity supplied by generators falls from Q^{14}_{G} in Figure 5-5 to Q^{24}_{G}; more capacity is supplied by demand response participants (Q^{24}_{DR} from Figure 5-6 vs. Q^{14}_{DR} from Figure 5-5).

The social deadweight loss avoided through demand response capacity supply is again the area shaded in green in Figure 5-6, but this area differs from that shaded area in Figure 5-5 in one critically important aspect. Since the market could not clear at any finite price without the demand response capacity program, the area of deadweight loss avoided through the demand response program is without limit at prices above P_{VG}--the price at which the supply curve for generation capacity become vertical. The arrow pointing upward indicates that the green shaded is unbounded, and therefore the value in terms of social value is very large. The challenge is to

---

[27] These conditions obtain because of the joint one-to-one relationship in supply of DR capacity and load reduction—that is, supplying capacity as a stock resource for a short time by reducing demand for a flow of electric energy. However, if generator capacity and capacity supplied by DR participants are not perfect substitutes from an electrical systems perspective, they may well be of different value to the market. This issue is addressed in the discussion in sections below.

find the appropriate high price that effectively places a finite limit on the size of the deadweight loss mitigated. The question is: should this be set at a value that reflects the cost of outages to consumers, such a value of lost load (VOLL), or some other large number?

Similarly, one can also examine the effect of the supply of capacity from demand response for the situation represented in Figure 5-3, where demand is above maximum capacity of generators due to an unforeseen outward shift in demand. This is depicted in Figure 5-7. The market again will *not* clear at any price in this situation, unless customers are allowed to bid capacity into the market through a demand response program. Here, the market will clear at a finite price of P^{34}_{G+DR} in Figure 5-7. Because the supply of capacity by generators is exhausted at prices above P_{VG}, and demand has shifted to the right, this equilibrium price is above P^{14}_{G+DR} from Figure 5-5. The capacity supplied by generators increases from Q^{14}_{G} in Figure 5-5 to Q^{34}_{G}. In this example, more capacity is also supplied by demand response participants (Q^{34}_{DR} from Figure 7 vs. Q^{14}_{DR} from Figure 5-5), but this may well depend on the size of the shift in demand and the nature of the supply curve of demand response capacity.

The social deadweight loss avoided through demand response capacity supply is again the area shaded in green in Figure 5-7, and this area differs from that shaded area in Figure 5-5 in the same critical aspect as that in Figure 5-6. Since after the demand shock the market could not clear at any finite price without the demand response capacity program, the area of deadweight loss avoided through the demand response program is without limit at prices above P_{VG}--the price at which the supply curve for generation capacity becomes vertical. This presents a similar challenge: to find the appropriate high price that effectively places a finite limit on the size of the deadweight loss mitigated. Should this be set at VOLL, or some other large number?[28]

Summary

In this section, an economic model was developed to provide a better understanding of the effects of allowing consumers as demand response participants to supply capacity and compete directly with generators in doing so. Although very stylized, the model offers important insights into these effects. Viewed as an additional source of supply of capacity, it should be no surprise that in a new equilibrium the price of capacity falls relative to what it would be otherwise, and generators supply something less than the full amount of the capacity requirement. The lower price for capacity may lead to reduced investment in new capacity in the future, but this is as it should be.

Customers are now able to help determine the value of additional capacity. If they are willing to self-supply some capacity through load reduction, then the need for future investment in generation capacity is reduced. Much of this result turns on the recognition that capacity supplied by demand response program participants is a jointly produced commodity, along with load reduction, and the joint production is in a one-to-one fixed proportion. Under normal system operation, these new supply resources can mitigate social deadweight losses in efficiency relative to the situation in customers are willing to supply capacity, but are not allowed to for administrative reasons.

---

[28] The size of the deadweight losses avoided through implementation of a DR capacity program will differ depending on the nature of the willingness of customers to reduce load. This is illustrated for two other DR supply curves for capacity in Appendix B.

It is also clear from this framework that while they are substitutes, for capacity provided through demand response programs is not a completely identical substitute for capacity from generators. As seen above, it is the joint nature of the provision of capacity and load reduction that makes demand response supplied capacity more fungible than capacity supplied by generators. It can be supplied in more divisible amounts, and if this capacity is under direct control of the system operator, it can be used to mitigate damages when the system is faced with unforeseen circumstances such as generator outages and abrupt changes in demand. By virtue of its joint production with load reduction, these resources can be used effectively to set priorities for service interruptions to avoid the need to interrupt customers whose value for load is very high. In this sense, there is also an implicit insurance or option value to demand response supplied capacity. One would need to be able to estimate the probabilities for these outcomes in order to estimate the expected insurance or option value of these demand response capacity resources. The value of these resources viewed from this perspective is in contrast to their value under normal circumstances where there is ample generation, but in which demand response supplied capacity merely serves as a substitute for generator capacity because customers are willing to supply it at a lower cost.

To shed further light on the value of demand response capacity resources, it is important to emphasize that the stylized markets for capacity presented here are most representative of single types of customers with a certain willingness to supply capacity. Three examples are highlighted, one in the text, and two more in Appendix B. However, at the market level the demand response supply curve for capacity will encompass the entire range in the behavior that could be exhibited by a heterogeneous mix of customers in any electricity market. And, in the aggregate, the market supply curve for capacity by demand response customers would be the horizontal summation of the supplies by the diverse set of customers. In the aggregate, this supply curve for capacity is likely to appear more as a continuous function with a positive slope when measured from right to left on a figure depicting a market for capacity. Alternatively, because of the one-to-one joint production of capacity and load reduction, movements along this supply curve from left to right would also trace out a continuous downward sloping demand curve for capacity. It is this type of aggregate supply/demand curve for capacity that is used in Section 6 to examine the effects of demand response supplied capacity in an ISO/RTO run capacity market. It is only at this level that one can understand the effects of allowing demand response participants to offer capacity when there is also an administratively-set downward sloping demand curve for capacity.

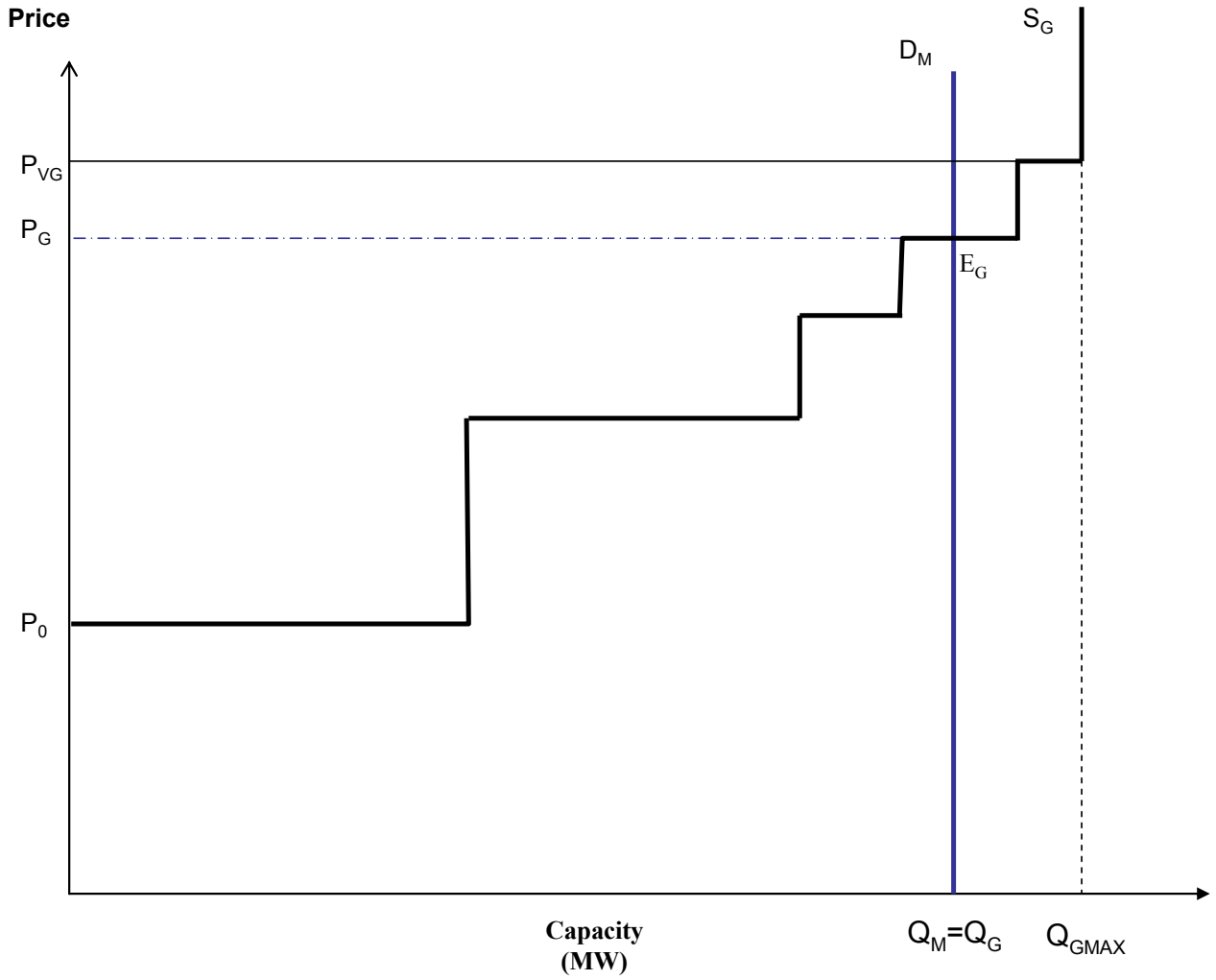


Figure 5-1

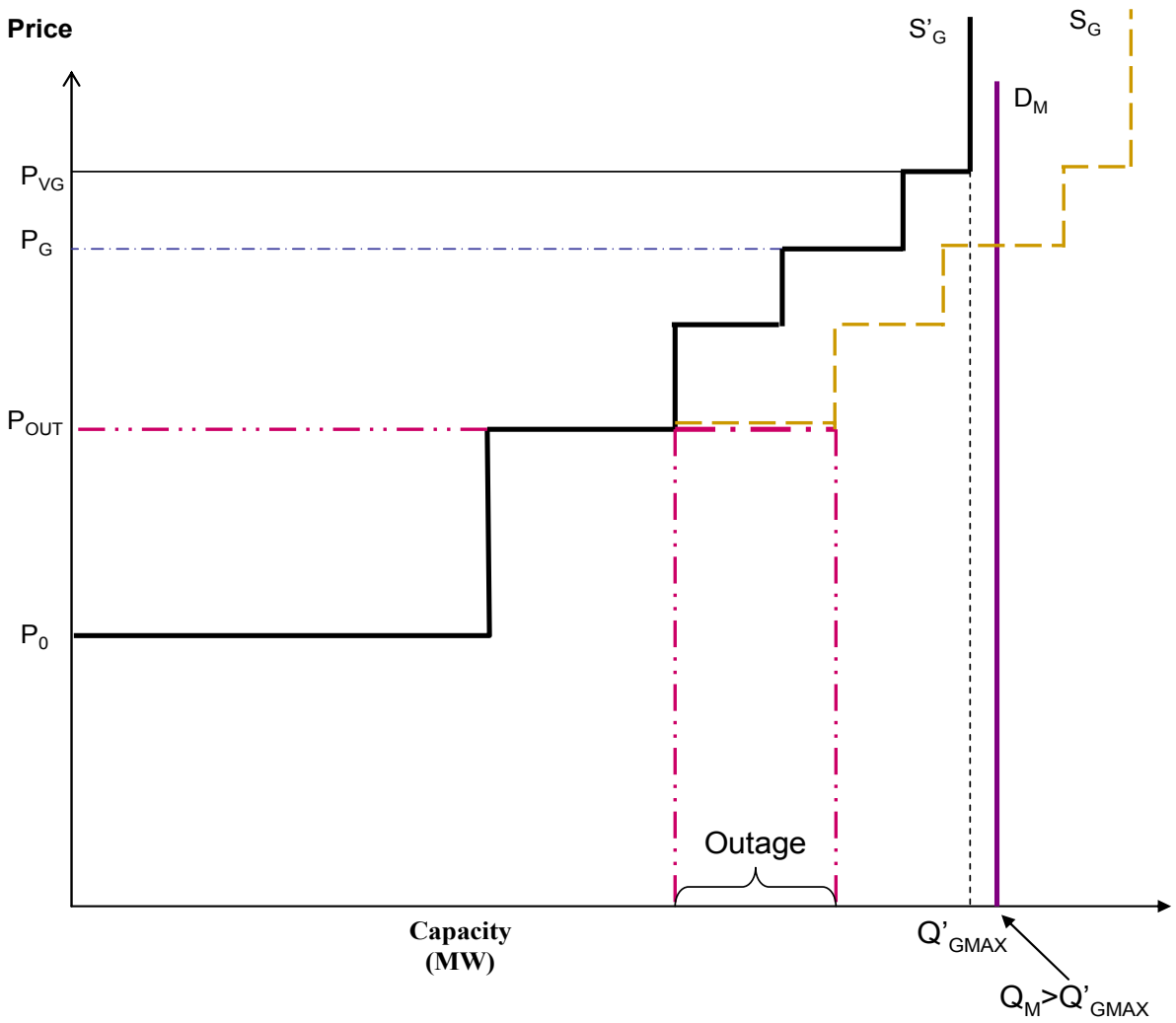

Price
S'G
SG
DM
PVG
PG
POUT
P0
Outage
Capacity
(MW)
Q'GMAX
QM>Q'GMAX

Figure 5-2

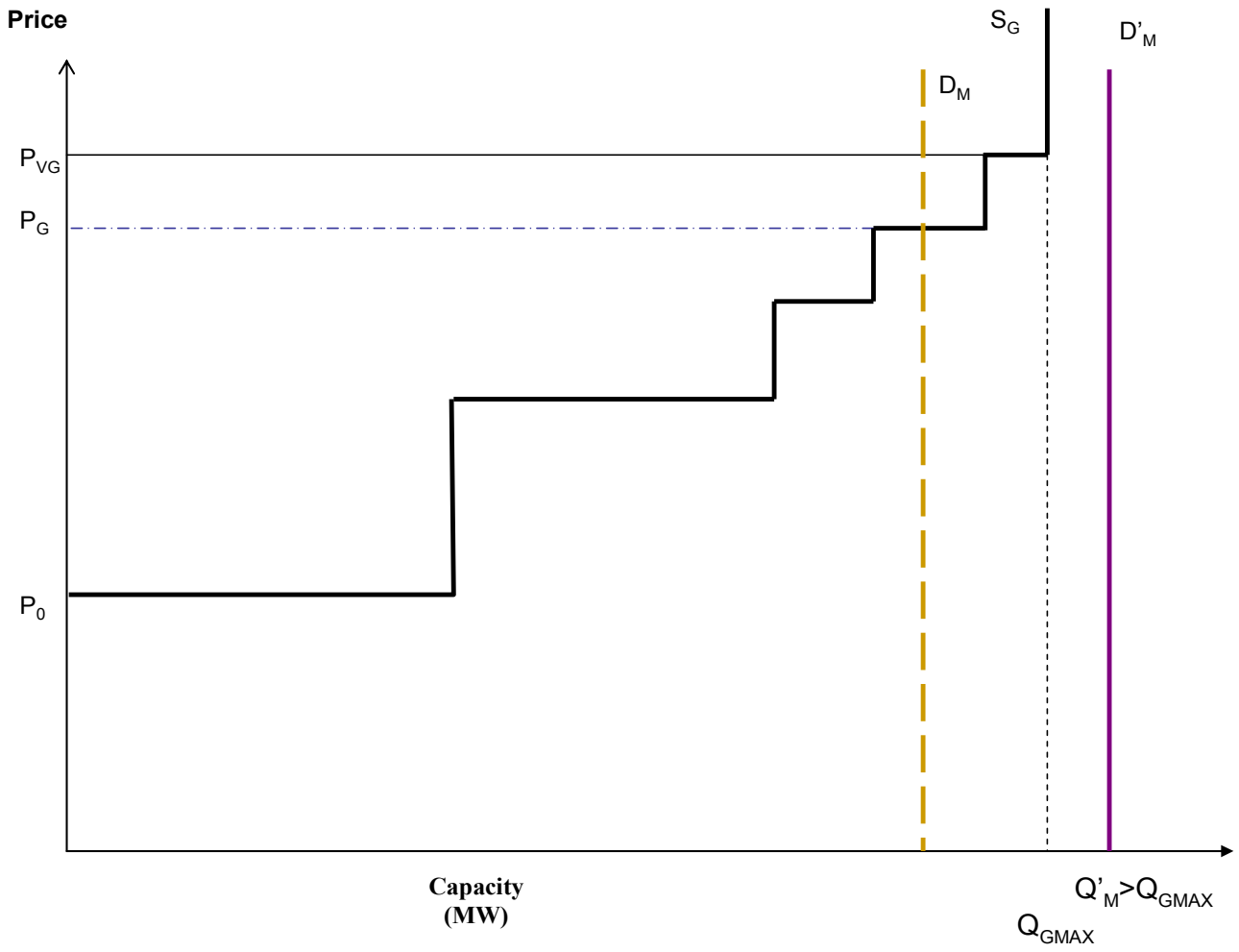


Figure 5-3

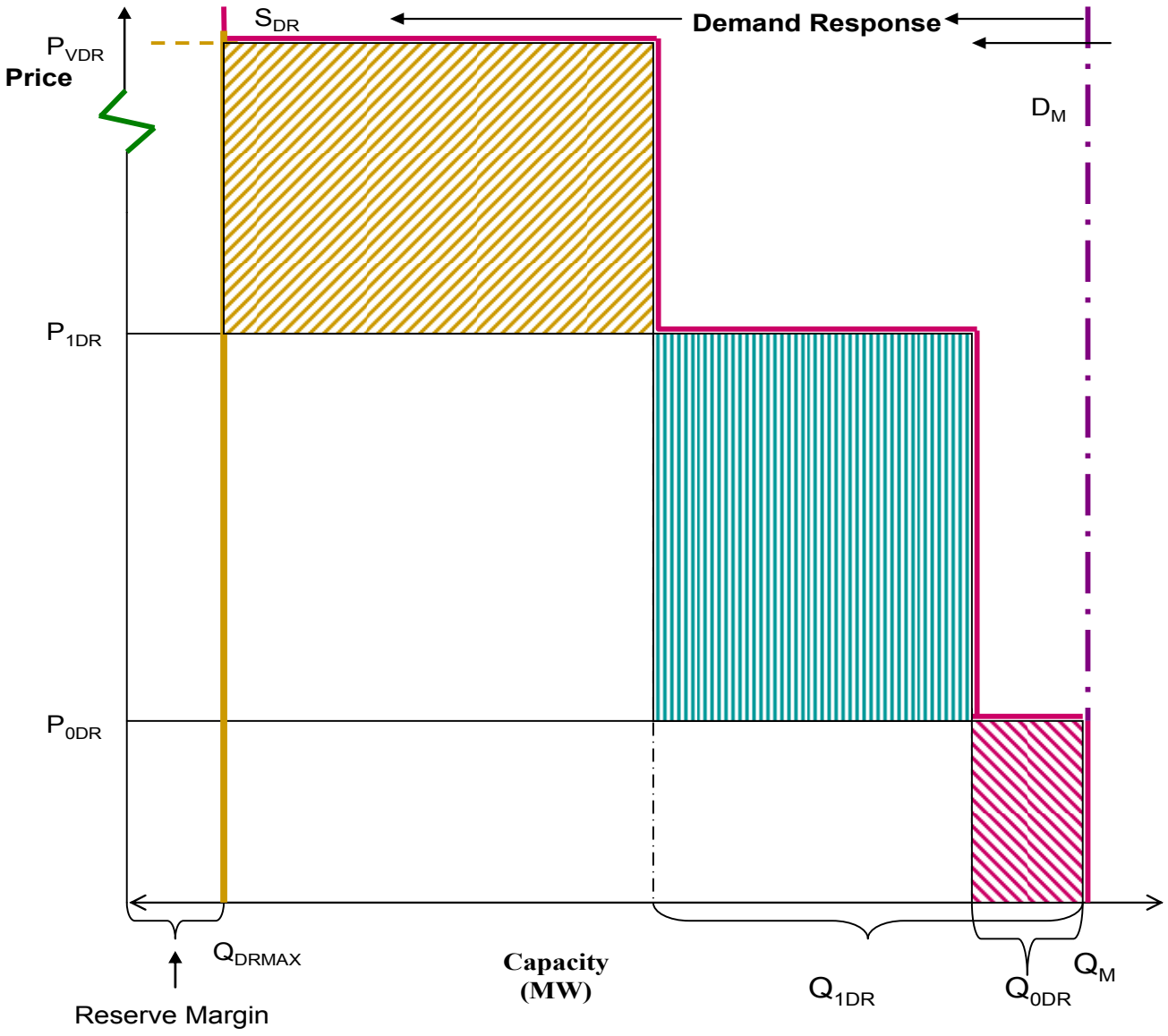

S_{DR}
Demand Response
P_{VDR}
Price
D_M
P_{1DR}
P_{0DR}
Q_{DRMAX}
Reserve Margin
Capacity
(MW)
Q_{1DR}
Q_{0DR}
Q_M

Figure 5-4

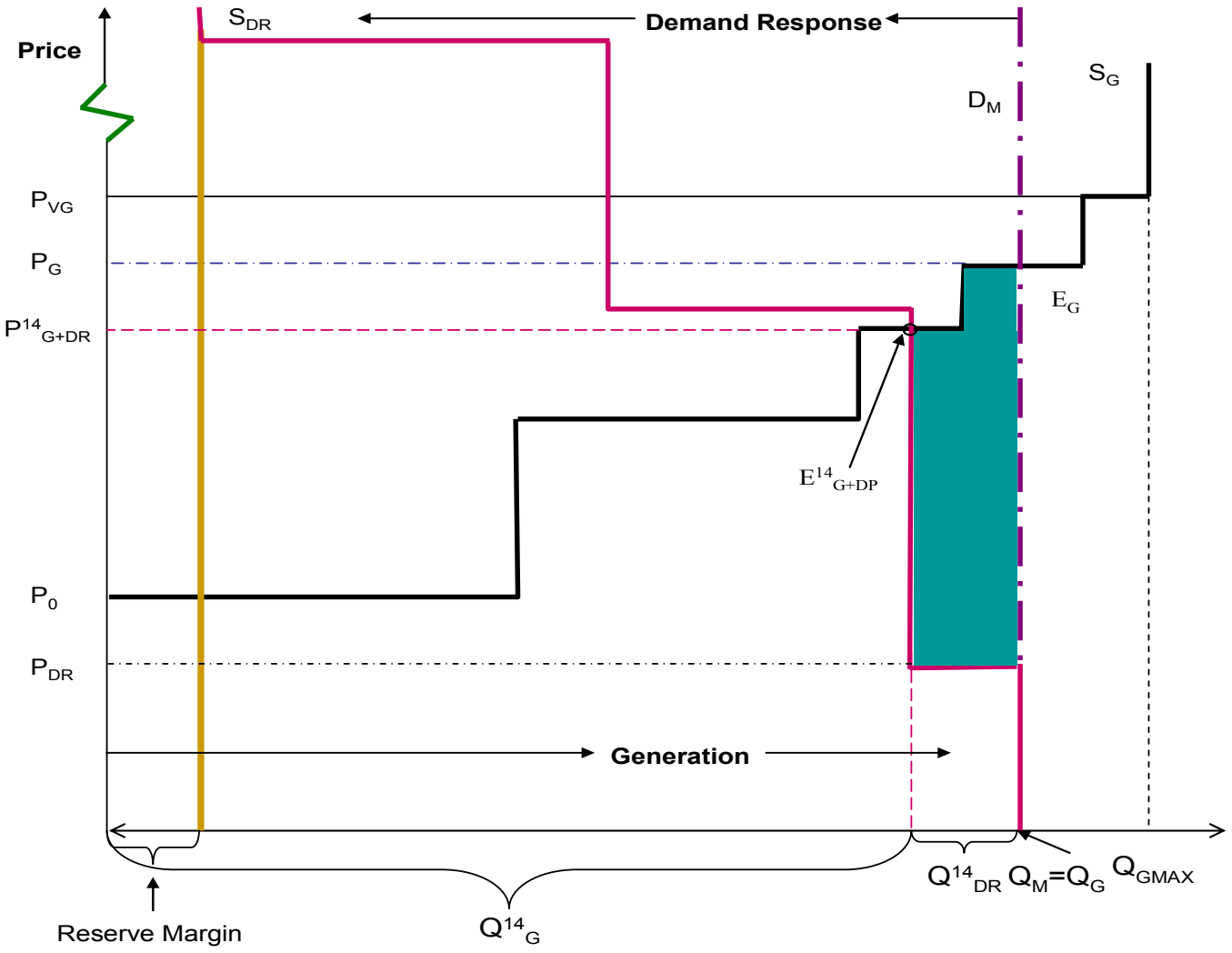

Price
S_{DR}
Demand Response
S_G
D_M
P_{VG}
P_G
P^{14}_{G+DR}
E_G
E^{14}_{G+DP}
P_0
P_{DR}
Generation
Q^{14}_{DR} Q_M=Q_G Q_{GMAX}
Reserve Margin
Q^{14}_G

Figure 5-5

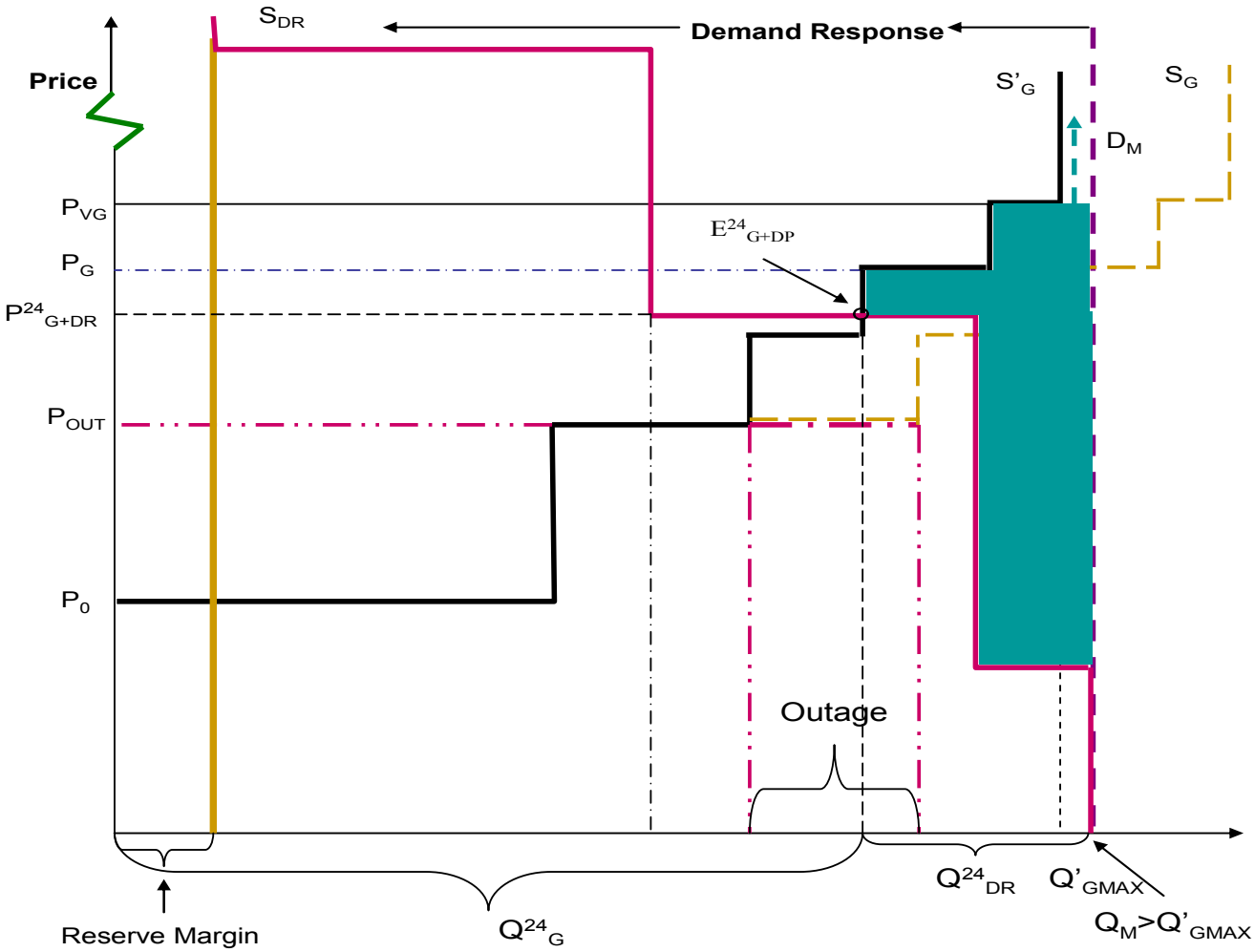

S_DR
Demand Response
Price
S'_G
S_G
D_M
P_VG
E^24_G+DP
P_G
P^24_G+DR
P_OUT
P_0
Outage
Q^24_DR
Q'_GMAX
Reserve Margin
Q^24_G
Q_M>Q'_GMAX

Figure 5-6

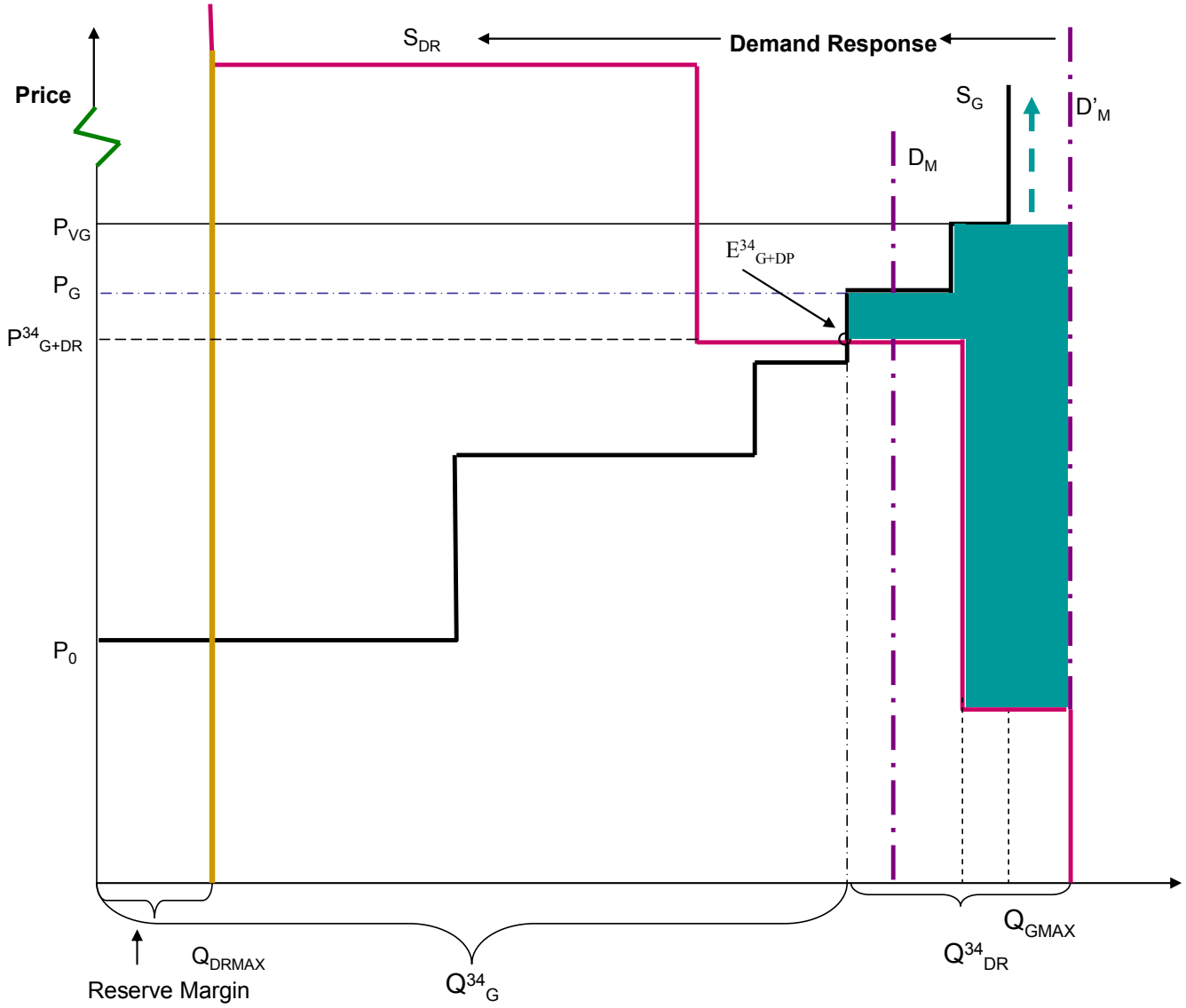

Price
S_{DR}
Demand Response
S_G
D'_M
D_M
P_{VG}
P_G
P^{34}_{G+DR}
E^{34}_{G+DP}
P_0
Q_{DRMAX}
Reserve Margin
Q^{34}_G
Q_{GMAX}
Q^{34}_{DR}

Figure 5-7

6
DEMAND RESPONSE AS A RESOURCE IN CENTRALIZED CAPACITY MARKETS

In the previous section a market was described whereby consumers select how much firm capacity to buy based on prevailing prices. While this framework depicts the demand for capacity by individual consumers, it does provide a basis of comparison for the current performance of existing centralized capacity markets.

To make these comparisons, it is necessary to understand how demand response impacts capacity markets where the demand response capability is treated as an equivalent to generation resources in meeting the reserve adequacy requirement. These impacts must be compared for two different circumstances. Under the first set of circumstances, it is assumed that the centralized procurement is for a specified and fixed amount of capacity. Second, the exposition is extended to markets in which the amount of capacity acquired is also determined administratively but under conditions in which the administrative authority characterizes a system-wide downward sloping demand for capacity.

To analyze this latter situation, it is convenient (and not limiting) to depict supply and demand for reserves as continuous (or nearly so) functions of price. In so doing, one can think of these market representations as the horizontal aggregation of the individual supply and demand curves for reserves by a number of generators and heterogeneous customer classes.

A Centralized Capacity Market with Fixed Demand for Capacity

A characterization of a centralized market for capacity is provided in Figure 6-1. It is similar to the situation depicted in Figure 5-1, except the supply of capacity by generators is assumed to be a continuous function of price at prices below $P_{VG.}$ At prices above P_{VG}, the capacity offered by generators equals all available generation capacity, and the supply curve becomes vertical. There is no more generation to be supplied at any price.

In the market depicted in Figure 6-1, the demand for generation is fixed at D_M, and the market clears at point E_G. The market clearing price is P^F_G, and this is at a quantity where the capacity supplied by generators is equal to the demand, at $S_G = D_M$. The amount of capacity supplied by generators is given by the distance Q^F_G. In the event that there no customers are willing to supply capacity through demand response programs, this equilibrium maximizes social welfare, provided that the administrative authority has set the level of capacity adequacy (including the 12 to 18% reserve margin) to reflect both the private and "public good" or social value of system-wide capacity.

Supply Curve for Capacity from Demand Response Resources

If, on the other hand, there are customers willing to reduce their demand for firm power at specified prices and provide capacity instead, the point E_G will not be a socially optimal equilibrium, even where there is a fixed demand for capacity. Put differently, due to the joint nature of supply of load reduction and capacity by customers, this situation gives rise to an

additional source of supply for capacity. These circumstances are depicted in Figure 6-2, where the aggregate supply of capacity by participants in a demand response program for system capacity is given by the curve labeled S_{DR}. Thus, the amount of capacity supplied by customers is again measured by moving from the right to the left in the figure. For example, at prices below P_{0DR}, participants in the demand response program for capacity are unwilling to supply any capacity. At a price of P_G, the amount of capacity supplied by demand response program participants would be the distance Q^F_{DR}.

In the event that these demand response providers are allowed to participate in the centralized capacity market, the new equilibrium capacity supply price will fall from P^F_G to P^F_{G+DR}. The total amount of capacity supplied to the system will still be at D_M, the level of capacity fixed administratively. But, because of demand response program participation, the amount of capacity supplied by generators will fall, from $Q_G = D_M$ in Figure 6-1 to Q^{FDR}_G in Figure 6-2. The difference between the fixed demand and this lower amount supplied by generators is now supplied by demand response program participants (e.g. the distance Q^F_{DR}). Since demand response participation in the centralized capacity market essentially shifts the supply curve for capacity outward at any price above P_{0DR}, the new equilibrium price in this market will fall to P^F_{G+DR}, despite the fact that the demand for capacity remains unchanged. This reduced value for capacity reflects the fact that some consumers are willing to forgo some level of firm service and supply needed capacity. They receive the market-clearing price for doing so, and other benefits from avoiding the higher price needed to ensure a higher level of future investment in new generation capacity.

The Implications for Economic Efficiency

The implications of this new source of supply of capacity by demand response participants are similar to those in Section 5. By being able to reveal their willingness to supply capacity at some price, and act accordingly, customers in demand response programs are able to reduce or eliminate the social deadweight losses in this market for capacity. The deadweight losses avoided are equal to the cross-hatched area in Figure 6-2. Although not demonstrated graphically, the demand response capacity resources remain more fungible than generator capacity and have insurance or option values in terms when unforeseen circumstances could compromise the reliability of the system if only generators can supply capacity. These insurance and option values are identical to those described in Section 5.

A Centralized Capacity Market with an Administratively Set Downward Sloping Demand for Capacity

It is clear from the above discussion that there are inefficiencies in the capacity markets if customers willing to participate in a demand response capacity program are not able to do so. Another consequence of this market where only generators can supply capacity is the significant variability in the price of capacity that comes from small shifts in demand along a supply curve that becomes very steep near maximum generation capacity. As in other markets where both demand and supply curves are inelastic (e.g. steep negative and positive slop, respectively), this volatility (e.g. uncertainty) in prices is likely to work against investment in new generation capacity.

In order to promote greater price stability, and to address concerns about market power and to provide a more stable revenue stream to resources, the NYISO and ISO-NE have incorporated an

administratively set downward sloping demand curve into their centralized capacity market designs (California Public Utilities Commission 2005) . While the introduction of this downward sloping demand curve for capacity may promote these objectives (Cramton and Stoft 2005), the challenge is to determine how this demand curve interacts with the supplies of capacity by generators and by participants in demand response programs to determine prices and market efficiency. The fact that demand response entails a one-to-one correspondence between joint production of load reduction and capacity lies at the heart of this assessment.

The Nature of Price Volatility in the Market for Capacity

To begin the analysis, it is instructive to explore further the market circumstances that give rise to extreme volatility in capacity prices. In so doing, one can better understand the rationale for characterizing administratively set capacity demand as price responsive. These circumstances are illustrated in Figure 6-3, where the initial equilibrium is again at E_G, the intersection between the generator's supply curve for capacity, S_G, and the fixed demand, D_M. To facilitate comparison, this is the same initial market equilibrium as in Figure 6-1.

The potential price volatility is represented in Figure 6-3 through an examination of the interaction between four situations. These situations are those in which for some unforeseen circumstances demand either falls, (e.g. shifts to D^-_M) or expands (e.g. shifts to D^+_M), and/or there is a shift in the generators' supply curve (e.g. an increase in supply, a shift to the right--to S^+_G or a decrease in supply, a shift to the left—to S^-_G).

There would clearly be a change in the market equilibrium (e.g. a change in the market clearing price and quantity of capacity) were any of these four situations to occur in isolation. However, the potential price volatility is captured best by two extreme situations in which unforeseen circumstances give rise to a simultaneous shift is supply of capacity from generators and a shift in demand. It is under these conditions that the market would see the most dramatic swings in the price of capacity.

The first of these two circumstances is when an increase in demand (a shift to the right) is accompanied by a shift to the left (reduced supply availability) in the supply curve of capacity by generators. This might be the result of an upsurge in the economy that expands forecasted demand concomitant with the retirement of older generation units that are no longer profitable to operate.

As seen in Figure 6-3, the difference in the equilibrium prices between these two sets of circumstance is large indeed—P^{F+}_{G-} vs. P^{F-}_{G+}. Because these demand curves are vertical (e.g. perfectly price inelastic) and the supply curves are very steep at this level of capacity (e.g. extremely price inelastic), the large change in price between these two extreme market equilibriums is accompanied by a very modest change in the equilibrium quantities of capacity—D^-_M vs. D^+_M. While there is still some volatility in price when any of these four unforeseen circumstances occurs in isolation, both the changes in price and capacity are somewhat more modest. To avoid too much undue clutter, these more modest changes in equilibrium prices and quantities are not depicted explicitly on Figure 6-3.

The Effects of an Administratively-set Demand Curve on Price Volatility

The effects of the administratively-set downward sloping demand curve for reserves on this price volatility are depicted in Figure 6-4. In this figure, the downward sloping demand curve for

capacity is designed to mimic that implemented by the NYISO in 2003. That is, demand is vertical for an amount of capacity well less than the required amount. Then, at some high price, indicated in Figure 6-4 as P_{AMAX}, the demand curve become horizontal, indicating that the NYISO will buy any amount of capacity from 0 to an amount D_{PAMAX}. However, for quantities of reserves greater than this amount, the price falls as one moves along the downward sloping part of the demand curve. To facilitate comparisons with the analyses above, the downward sloping demand curve is positioned to intersect the original generators' supply curve for capacity at a point where demand is equal $D_M = S_G$, the original level of fixed level of demand from Figure 6-1. It is reasonable to assume that this equilibrium is where generators' supply of capacity equals required capacity (peak load plus the reserve margin).

Since this demand curve slopes downward, it now remains in a fixed position. But circumstances can still give rise to shifts in the supply schedule for generation capacity. However, due to the administratively-set demand curve, the difference in price due to these shifts in supply by generators (and absent any shifts in a vertical demand schedule) is attenuated; it is now P^S_{G-} vs. P^S_{G+}. This range is much smaller that than depicted under the market conditions reflected in Figure 6-3. Yet, the difference in equilibrium quantities of capacity, D^{S-}_M vs. D^{S+}_M, is of similar magnitude.

Absent any opportunity (or willingness) on the part of customers to supply capacity through demand response programs, one might well conclude that this administratively set demand curve is an effective mechanism to mitigate price volatility. But, for the market equilibrium to be an efficient one, this administratively-set demand curve must also be correctly positioned near the target reserve margin so that equilibrium prices reflect the normal cost of new generation or the annualized fixed cost of a benchmark generator.

Supply of Capacity from Demand Response Resources

When customers are also allowed to compete in this capacity market through a demand response program, the situation is more complex. As indicated above, this complexity arises due to the joint nature of customers' supply of capacity and load reduction. In Figure 6-5, the supply of capacity by demand response customers is given by the curve S_{DR}, but the supply is measured from right to left. The supply curve begins where the supply curve intersects the administratively determined demand curve at a price of P_{0DR}—the price below which no capacity is supplied by demand response customers.

Because of the joint nature of production, it is also true that if one moves down the curve S_{DR} from left to right, it reflects the consumers' demand for capacity. Thus, there are now two demand curves for capacity in Figure 6-5. If demand response customers can participate in the market, their demand curve is the operable one, provided that some demand response program participants are willing to supply capacity (by dropping load) at prices below P^S_G, the equilibrium price when only generators supply capacity.[29]

When this demand response, jointly determined demand/supply curve for capacity is controlling, the new equilibrium is at E^S_{G+DR}--the intersection of S_{DR} and S_G, as it is in Figure 6-2. The total

[29] Of course, in the case where demand response program participants are unwilling to supply and capacity at prices below this equilibrium price, the administratively-set demand curve remains the operative one, and the implementation of a demand response program for capacity would have no effect on the capacity market.

supply of capacity is now equal to that supplied by generators, Q^{SDR}_{G}, plus that supplied by demand response participants, indicated as the bracketed quantity Q^{S}_{DR} in Figure 6-5. The new equilibrium price is P^{S}_{G+DR}, slightly higher than that in Figure 6-5 of P^{F}_{G+DR}. This is as it should be because of the increased administrative demand for capacity (over and above the demand by customers—as reflected in the left to right movements along the supply curve of demand response capacity) at prices below P^{S}_{G}.

The Implications for Economic Efficiency

Accordingly, this administratively-set downward sloping demand curve for capacity leads to higher social dead weight losses (the shaded area in Figure 6-5) compared with the situation where the demand for capacity is fixed demand. The only case where this would *not* be true is if the administratively-set demand curve by chance would coincide with the supply curve for capacity by demand response participants.

Because customers' willingness to supply capacity differs at different prices, there is no guarantee that this supply schedule, when offered in competition with the supply from generators, would lead to an equilibrium price at or near the cost of new generation. Indeed, the price might well be below this level because of customers' willingness to self-supply capacity through load reduction. This lower equilibrium price would send an appropriate signal to potential investors in new capacity: When there is capacity supplied through demand response there is a need for less new generation capacity in the future than there might be otherwise.

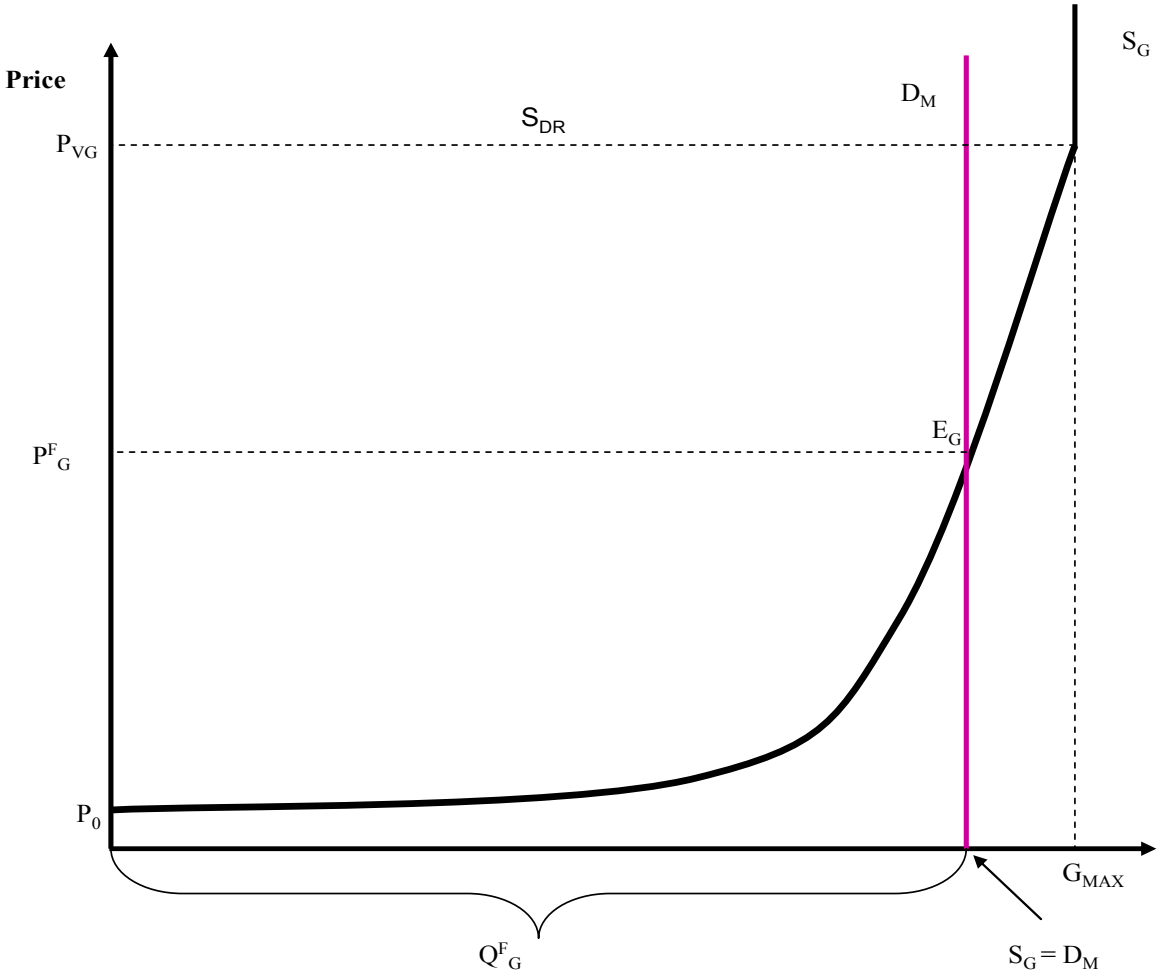


Figure 6-1

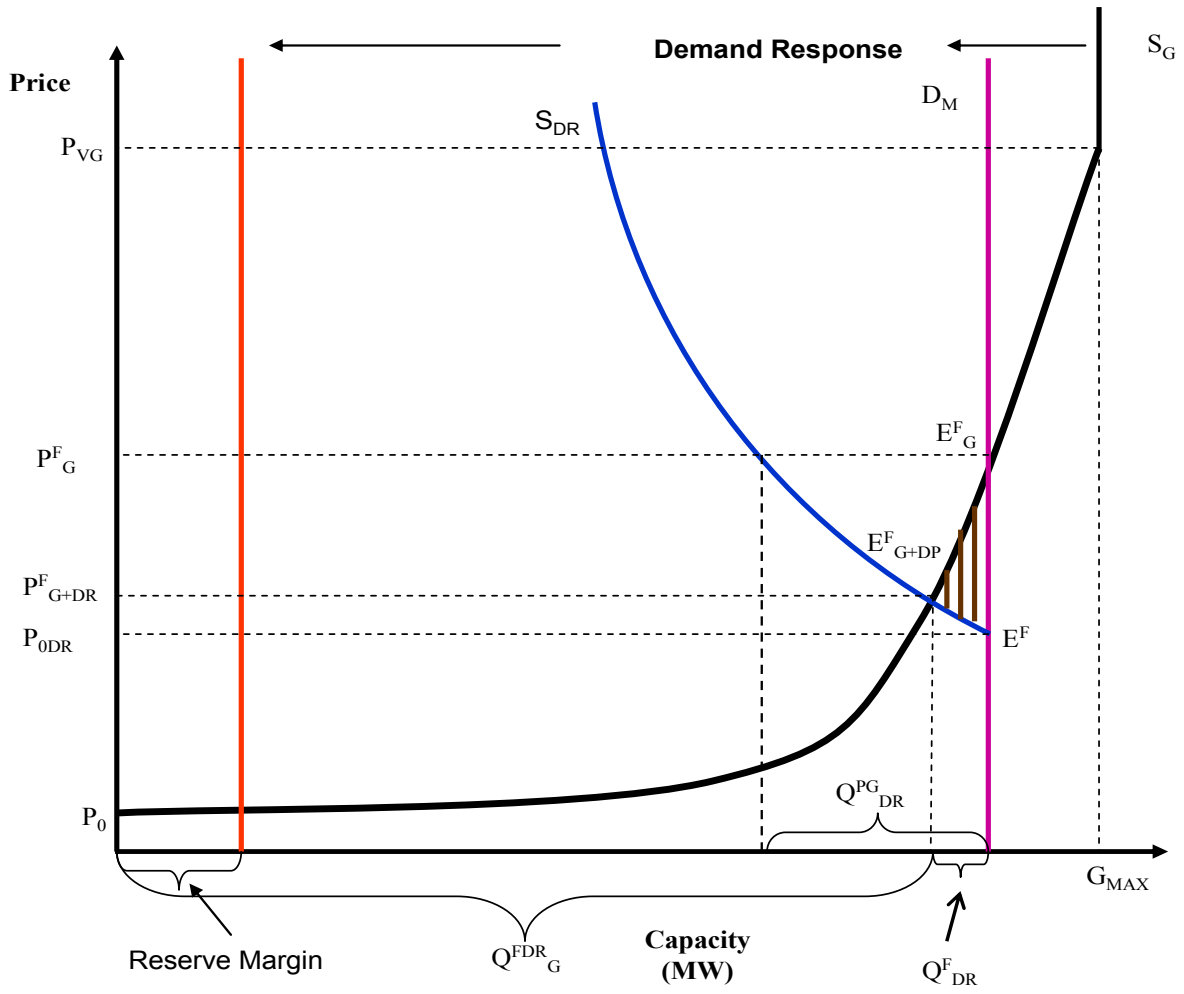

Demand Response
Price
S_G
D_M
S_{DR}
P_{VG}
P^F_G
E^F_G
E^F_{G+DP}
P^F_{G+DR}
P_{0DR}
E^F
P_0
Q^{PG}_{DR}
G_{MAX}
Reserve Margin
Q^{FDR}_G
Capacity
(MW)
Q^F_{DR}

Figure 6-2

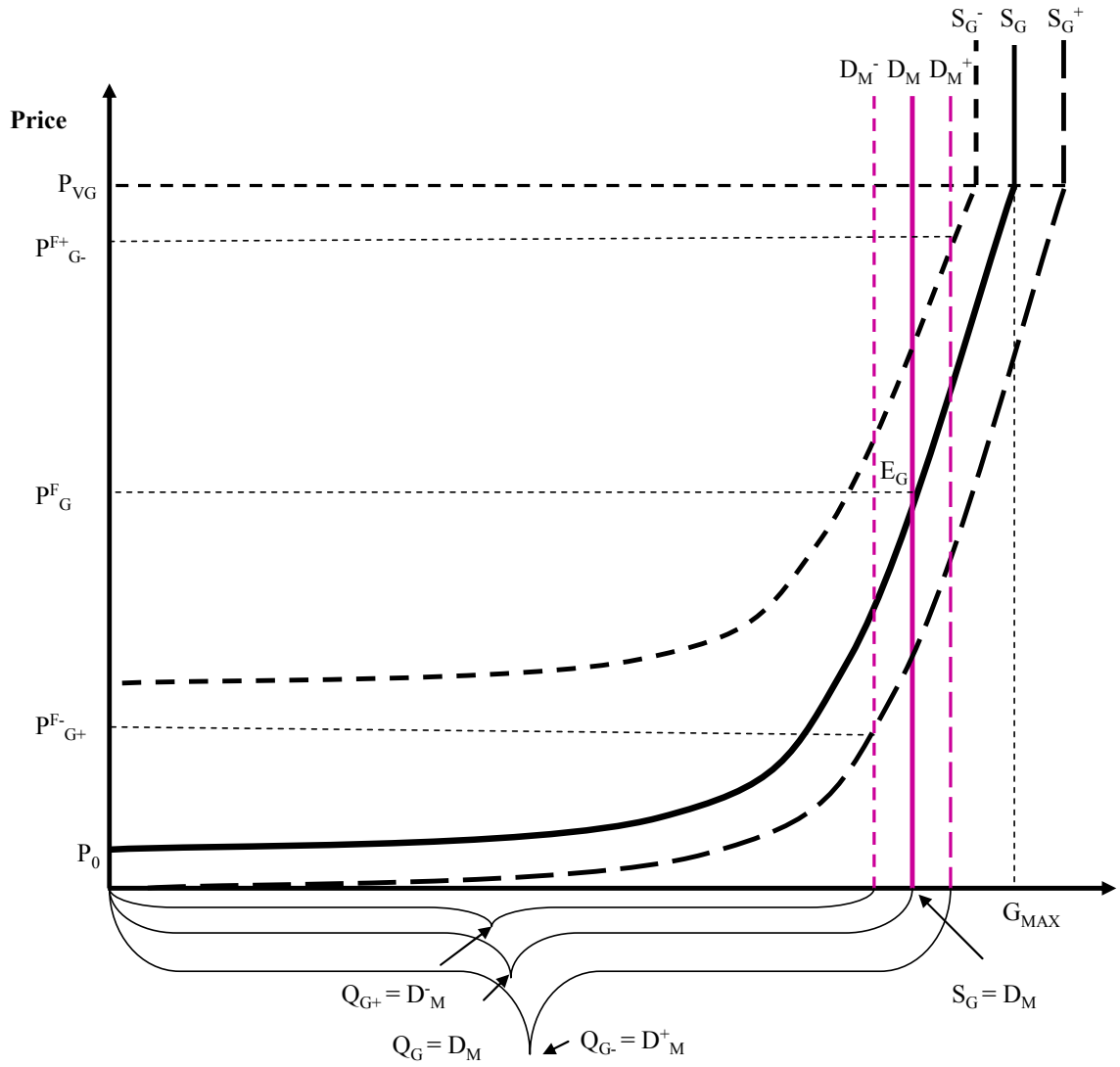

Price
S_G^- S_G S_G^+
D_M^- D_M D_M^+
P_{VG}
P^{F+}_{G-}
P^F_G
E_G
P^{F-}_{G+}
P_0
G_{MAX}
$Q_{G+} = D^-_M$
$Q_G = D_M$
$Q_{G-} = D^+_M$
$S_G = D_M$

Figure 6-3

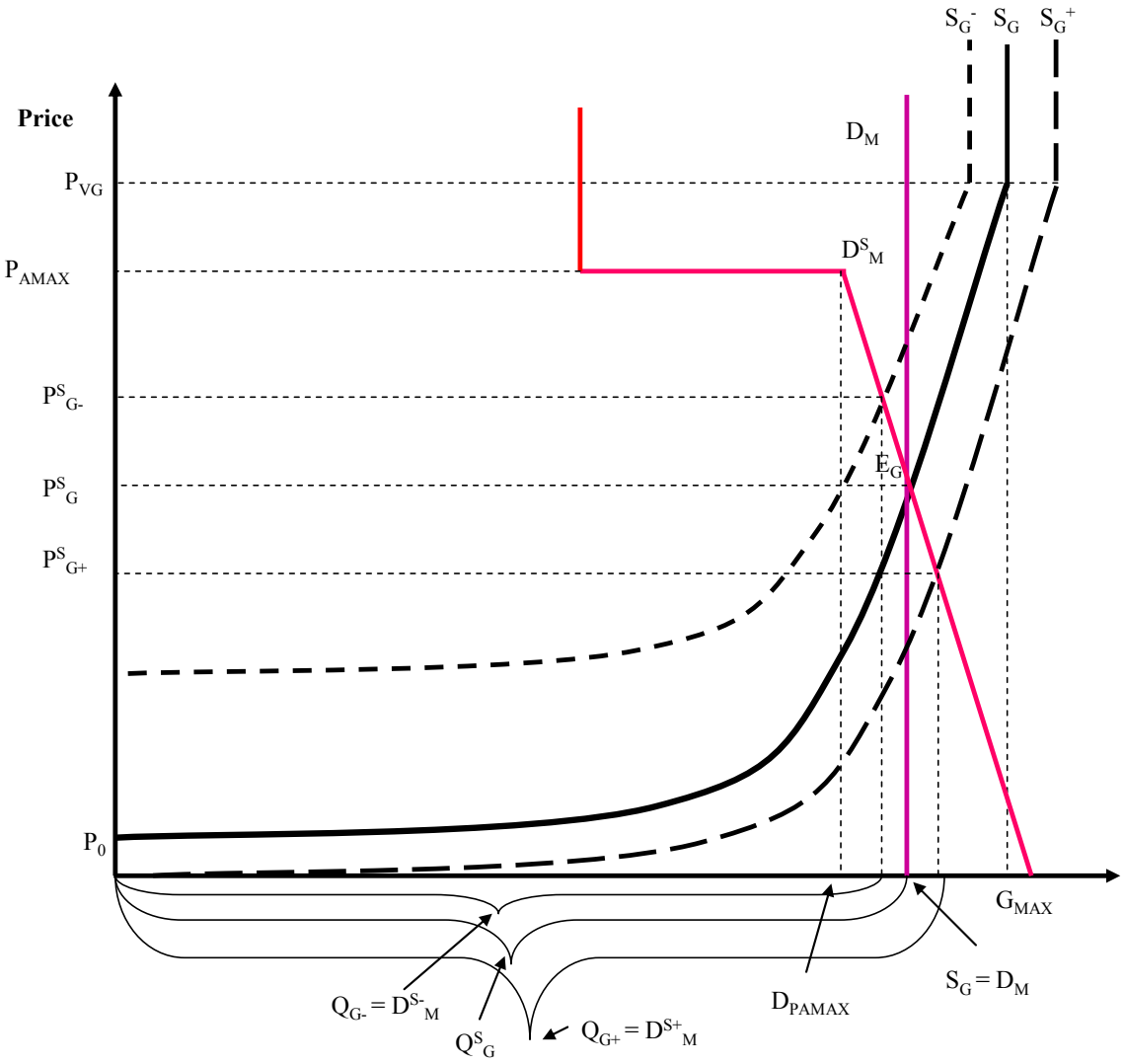

Price
S_G^-
S_G
S_G^+
D_M
P_{VG}
D^S_M
P_{AMAX}
P^S_{G-}
E_G
P^S_G
P^S_{G+}
P_0
G_{MAX}
$Q_{G-} = D^{S-}_M$
Q^S_G
$Q_{G+} = D^{S+}_M$
D_{PAMAX}
$S_G = D_M$

Figure 6-4

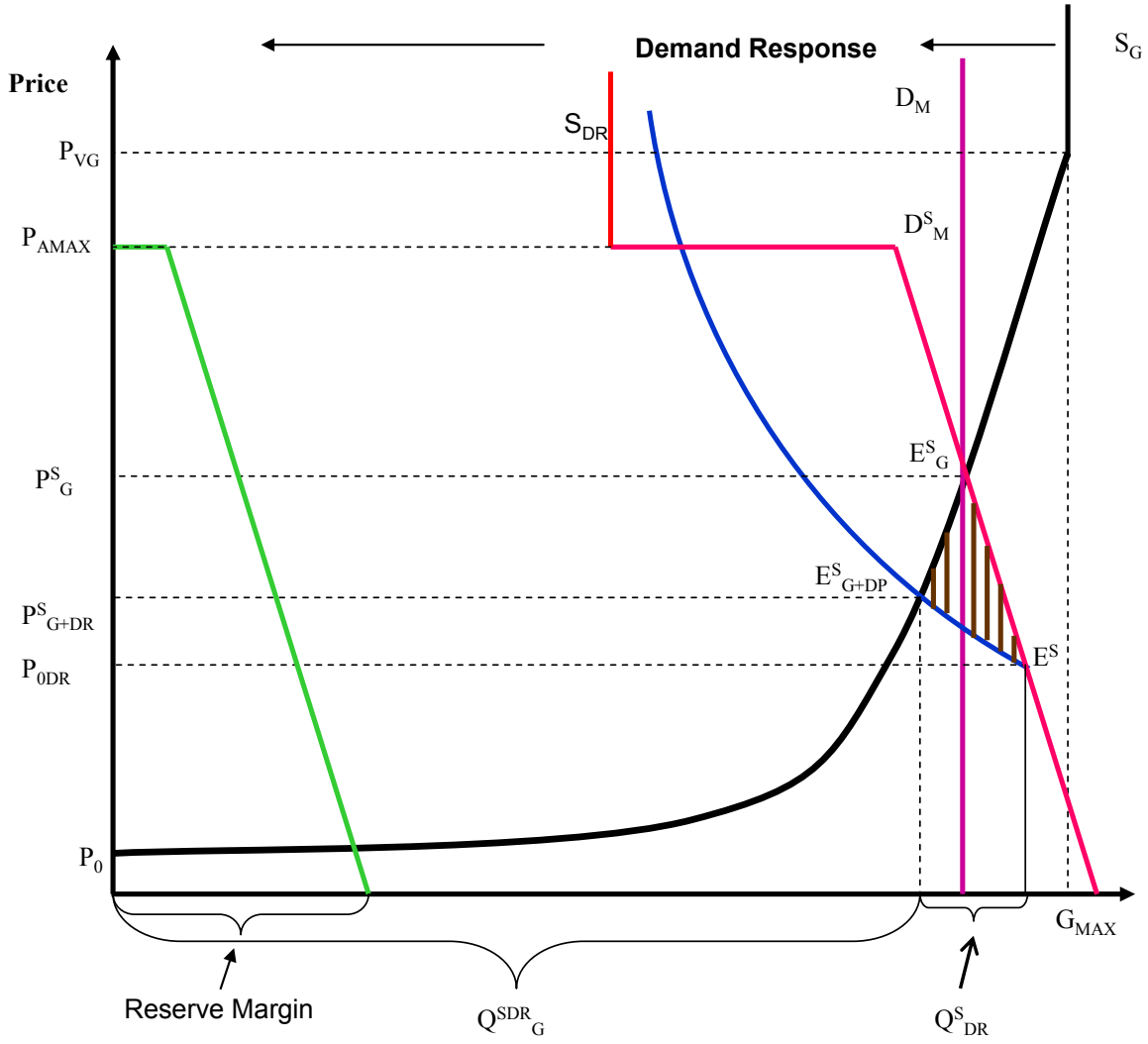

Demand Response
Price
S_G
S_{DR}
D_M
D^S_M
P_{VG}
P_{AMAX}
E^S_G
P^S_G
E^S_{G+DP}
P^S_{G+DR}
E^S
P_{0DR}
P_0
G_{MAX}
Reserve Margin
Q^{SDR}_G
Q^S_{DR}

Figure 6-5

7
DEMAND RESPONSE AS A RESOURCE FOR VERTICALLY INTEGRATED UTILITIES

The previous section described how demand response can influence the cost of capacity in ISO/RTO wholesale markets. Much of the country is served under less organized conditions. Either the ISO/RTO does not operate a capacity market, or vertically integrated utilities and cooperatives and MUNYS operate monopoly franchises, and they are responsible for meeting reliability obligations to electricity consumers.

To compare the value of capacity supplied through demand response in an ISO/RTO-run electricity market with its value in a vertically integrated utility, it is important to remember that these utilities operate in a market that has no centrally imposed capacity requirement. A vertically integrated utility must still arrange to fulfill capacity obligations to meet resource adequacy requirements set by local reliability councils (under NERC's coordination). These requirements are often enforced by a state or other regulatory body. Moreover, the recovery of costs associated with the acquisition of capacity assets is regulated, and this influences both when and what types of assets are acquired.

Without access to any organized market for capacity, vertically integrated utilities must establish those capacity requirements through introspective (as opposed to market-wide) analyses. These analyses characterize enterprise capacity resource availability and consider demand forecasts explicitly, and in so doing, the process involves many aspects of wholesale market prices. In the final analysis, the utility must meet these capacity requirements by either building its own capacity or by acquiring the capacity rights to generation plants owned by others.

However, because the planning perspective is that of the utility, the marginal value of capacity to the utility depends on the character and level of demand to be served, what resources are available, and how the utility dispatches them to meet its load obligations.[30] It is reasonable to assume that as a part of the planning perspective those involved in seeing that the capacity requirements are met have a good idea of how much capacity can be purchased at various prices, and these could be organized according to the least expensive to the most expensive (a sort of informal bid stack if you will) which is useful in making decisions about the order in which generation capacity is purchased. Utilities have considerable expertise at conducting comprehensive planning studies to ascertain what level and types of capacity are required to meet their reliability requirement. Capacity expansion models provide a detailed characterization of supply needs, which involve diverse technologies with different costs. Moreover, the utilities are experienced buyers of capacity, either by capitalizing and constructing units or by purchasing the capacity of merchant generation. In other words, in a utility's efforts to examine and place a

[30] The report by the U. S. Department of Energy (2006) describes the nature and character of the differences in how the value of demand response capacity is determined by a vertically integrated utility that holds a localized retail monopoly.

value on the capacity supplied by demand response participants, the utility will operate under conditions that are not too different from those that characterize organized capacity markets.

They can evaluate demand response directly or indirectly to ascertain the avoided cost attributable to demand response. The direct approach involves incorporating the effects of demand response into the capacity planning process. Demand response resources can be depicted as generation units that are dispatched to meet load. Alternatively, the load-modifying capabilities of a specified level (or levels) or demand response can be incorporated into an alterative demand scenario. Under either specification, the planning model is run with and without demand response. The difference in the capacity supply cost is the expected savings attributable to the demand response. This treatment of demand response in the planning model amounts to an empirical manifestation of the conceptual model used above to portray how demand response influences capacity prices in wholesale markets.

The indirect valuation method results from invoking a convenient (but limiting) assumption; demand response resources displace only peaking capacity units that operate infrequently. Therefore, their value is equal to the cost associated with maintaining such units in the generation portfolio. For a utility, that cost is the annualized carrying cost, which is expressed in terms of $/kW year. Under this construction, demand response resources can be paid up to that amount. Typically, the utility procures demand response resources from its customers through a tariff offering. If doing so results in additional costs such as recruitment, metering and settlement, then those costs are deducted to derive the avoided cost of demand response--the most that should be paid to participants.

Will the results be the same? For two markets that were equal in every way in terms of the character of supply and demand, one would expect relatively small differences in the capacity value of demand response that a market solution would achieve and what a utility would be willing pay. There are many reasons why this would not be the case. They include:[31]

- A market sets the value of capacity at the market-clearing price. All generation receives that price. A utility that uses avoided costs pegs the value of demand response to the carrying cost of a peaking unit. Depending on the stock of generation available in the market, the market-clearing price may be set by a peaking unit, or by some other, more or less expensive unit. One of the rationales for using the demand curve for capacity is that is imposes the cost of a peaking unit on to the market by virtue of way the demand curve is set--the downward sloping section begins at such a price.
- The size and technology of the marginal unit may not be the same due to localized conditions. In making capacity expansion decisions a utility often considers security needs. So, although the expansion is needed to serve peak demand, it might select a mid-range unit rather than a peaking unit to bolster its stock of generation that can provide both base load capability for part of the units' capacity and operating reserves from the upper range of its operating capability.
- Markets clear regularly, every year is the current practice, which results in a constant stream of capacity prices that reflect changes in supply. Prices will fluctuate, but along a trend line that reflects technology costs and investor expectations. A utility would make investments only routinely. Because of the indivisibilities of generation, there will be times when it has

[31] Neenan and Hemphill (2008) review utility capacity avoided cost methodologies.

surplus capacity, and its avoided costs will be zero. If demand response can contract and expand quickly, then it will help smooth out the impacts of large but only periodic generation investments.

- On the other hand, it may be preferable to take a longer term view of demand response by calculating the avoided cost over several years, and calculate a levelized cost that is offered to demand response each year.

There are other reasons for variations in the value of demand response across markets at any period in time, and over time, such as: the level and nature of demand growth, the influences of investments in energy efficiency by consumers, the age of the stock of generation, the fuels (and fuel prices) that are used to generate electricity, and the level and nature of the prices of electricity charged to consumers.

These are factors that shift the demand and supply curves that characterize a wholesale market. So, if they can be incorporated into an empirical representation of the market, their influences on market prices can be isolated and explained. Similarly, the effects of these influences on a utility's planning for capital expansion can be unbundled, at least to some extent. Efforts to determine the size of these effects on the prices and/or value of capacity are clearly worthwhile, but they are beyond the scope of this research effort.

8
FROM THEORY TO PRACTICE

A conceptual framework was developed to describe how capacity supply is influenced by programs that allow customers to express their demand for reliability. According to that framework, if demand for reliability is not completely inelastic, and instead at least some consumers would elect a level of reliability that is lower or higher than the universally imposed standard at prevailing supply costs, then the efficiency of electricity markets can be improved by implementing measures that reveal those preferences.

The benefits appear to be equally achievable by having consumers elect the level of reliability they want by nominating a firm supply and providing that level, or by allowing them to bid to supply capacity through demand response. However, an administratively fashioned capacity demand curve improves market performance only if it is in fact truly characteristic of underlying consumer preferences. This likely would be the case only by chance unless it was rigorously constructed from revealed or stated preferences.

The benefits associated with demand response as a capacity resource are reductions in deadweight efficiency losses. In the context of this framework, these losses are portrayed as areas between supply and demand curves in unspecified price and quantity space. This provides a characterization of the factors that determine the relative size of the benefits under alternative market supply and demand conditions. But this portrayal does not indicate the nominal level of the efficiency improvement that can be associated with demand response, nor the actual monetary benefits that would accrue to program participants.

The actual level of benefits is important to know because it serves as a measure against which the costs of implementing market changes should be compared. There may be instances where the gains are too modest to justify the cost.. Also, there are a number of ways in which market restructuring can be undertaken, and stakeholders should be know the net benefits when choosing among them.

The next step in this line of research is to develop an empirical version of the conceptual framework. This will involve the estimation of supply and demand equations that are structurally consistent with the conceptual model, but are flexible enough to account for the many factors that distinguish electricity markets.

A parallel inquiry is needed to identify the mechanisms that would be required to implement the alternative structures. In such an inquiry, one must determine how consumers would nominate firm power demands for a demand subscription service, and develop the protocols for effective supply of demand response as an alternative source of capacity. The latter can build on the experience of ISO/RTOs in integrating demand response into wholesale markets. The development of a system to allow consumers to nominate firm power levels presents greater challenges. Under both designs, protocols must be developed to establish how and when curtailments are called so that demand response capacity is a near perfect substitute for the generation capacity it displaces.

References

1. Barbose, G., C. Goldman, and B. Neenan (2004). "A Survey of Utility Experience with Real-Time Pricing," Lawrence Berkeley National Laboratory Report No. LBNL-54238. Available at http://www.lbl.gov/
2. Barbose, G., Goldman, C, and Neenan, B. (2006). "The Role of Demand Response in Default Service Pricing." *Electricity Journal*, 19 (April) pp. 64-74,
3. Baumol, W. and W. Oates (1988). *The Theory of Environmental Policy*. (2nd ed.) Cambridge: Cambridge University Press.
4. Beattie, B. C. R. Taylor and & M. Watts (2009). *The Economics of Production*, 2nd edition, Malabar, FL: Krieger Publishing Company.
5. Boisvert, R. and B. Neenan (2003). "Social Welfare Implications of Demand Response Programs in Competitive Electricity Markets," LBNL-52530 Environmental Energy Technologies Division Lawrence Berkeley National Laboratory, Berkeley, CA.
6. Borenstein, S. (2005a). "The Long-Run Efficiency of Real-Time Electricity Pricing," *The Energy Journal*, (26) pp. 93-116.
7. Borenstein, S. (2005b). "Time-Varying Retail Electricity Prices: Theory and Practice," in (eds.) J. Griffin, J. and S. Puller (eds.) *Electricity Deregulation: Choices and Challenges*, Chicago, IL: The University of Chicago Press, pp. 317-357.
8. California Public Utilities Commission (2005). "Capacity Markets White Paper," CPUC Energy Division, San Francisco, CA.
9. Cramton, P. and S. Stoft (2005). "A Capacity Market that Makes Sense," *Electricity Journal* 18(August/September) pp. 43-54.
10. Federal Energy Regulatory Commission (2009). "A National Assessment of Demand Response Potential", Staff Report prepared by the Brattle Group, Freeman, Sullivan & Co., and Global Energy Partners, LLC.
11. Field, B. (2001). *Natural Resource Economics: An Introduction*, New York: McGraw-Hill.
12. Goldman, C., N. Hopper, R. Bharvirkar, B. Neenan, and P. Cappers (2007). "A Methodology for Estimating Large-Customer Demand Response Market Potential, "Energy Analysis Department, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA.
13. Heffner, G. (2009). "Demand Response Valuation Frameworks Paper." Demand Response Research Center, Lawrence Berkeley Laboratory LBNL-2489E.
14. Hemphill, R. and B. Neenan (2008), "Characterizing and Quantifying the Societal Benefits Attributable to Smart Metering Investments," EPRI, Palo Alto, CA. 1017006.
15. Hobbs, B., J. Inon, and S. Stoft (2001). "Installed Capacity Requirements and Price Caps; Oil on the Water, Fuel, or Fire?" *Electricity Journal* 14(July) pp. 23-34.
16. ISO/RTO Council Markets Committee (2007a). "Harnessing the Power of Demand: How ISOs and RTOs are Integrating Demand Response into Wholesale Electricity Markets."
17. ISO/RTO Council Markets Committee (2007b). "North American Wholesale Electricity Demand Response Program Comparison."
18. Just, R., D. Hueth, and A. Schmitz (2004). *Applied Welfare Economics and Public Policy*, Northampton, MA: Edward Elgar.

19. Kirby, B., J. Dyer, C. Martinez, R. Shoureshi, R. Guttromson, and J. Dagle (2002). "Frequency Control Concerns in the North American Electric Power System," Oak Ridge National Laboratory Report ORNL/TM-2003/41. http://www.ornl.gov/
20. Kirby, B., and E. Hirst (2000). "Customer-specific Metrics for the Regulation and Load-following Ancillary Services," Oak Ridge National Laboratory Report ORNL/CON-474. http://www.ornl.gov/.
21. Litchenburg and D. Zilberman (1986). "The Econometrics of Damage Control: Why Specification Matters," *American Journal of Agricultural Economics*, 68(May) pp. 261-73.
22. Mas-Colell, A., M. Whinston, and J. Green (1995). *Microeconomic Theory*, New York: Oxford University Press.
23. Neenan, B, and J. Eorn (2008). "Price Elasticity of Demand for Electricity: A Primer and Synthesis", White Paper, EPRI, Palo Alto, CA. 10162642007.
24. Neenan, B; Hemphill, R. (2008). "Societal Benefits of Smart Metering Investments," *Electricity Journal*. 21(October) app.-45.
25. NERA (2007). "Independent Study to Establish Parameters of the ICAP Demand Curve for the New York Independent System Operator" NERA Economic Consulting, Washington, DC.
26. Oren, S. (2005). "Ensuring Generation Adequacy in Competitive Electricity Markets," in (eds.) J. Griffin, J. and S. Puller (eds.) *Electricity Deregulation: Choices and Challenges*, Chicago, IL: The University of Chicago Press, pp. 388-414.
27. Rohmund, I., G. Wikler, K. Smith, S. Yoshida, A. Faruqui, R. Hledik, and S. Sergici (2009). "Assessment of Achievable Potential from Energy Efficiency and Demand Response Programs in the U.S. (2010–2030)," Technical Report, EPRI, Palo Alto, CA. 1016987.
28. Ruff, L. E. (2002). "Economic Principles of Demand Response in Electricity", A Report to Edison Electric Institute. September 3.
29. Schwarz, P., T. Taylor, M. Birmingham, and S. Dardan (2002). "Industrial Response to Real-Time Prices for Electricity and Utilities". *Economic Inquiry*, forthcoming.
30. Stoft, S. (2002). *Power System Economics: Designing Markets for Electricity*, IEEE Press, Wiley-Interscience, New York: John Wiley & Sons, Inc.
31. Spulber, D. (1985). "Effluent Regulation and Long-Run Optimality". *Journal of Environmental Economics and Management* 12, pp. 103-116.
32. Spulber, D. (1985). Regulation and Markets, Cambridge, MA: The MIT Press.
33. Sullivan, M, M. Mercurio, and J. Schellenberg 2009. "Estimated Value of Service Reliability for Electric Utility Customers in the United States," Prepared for Office of Electricity Delivery and Energy Reliability, U.S. Department of Energy. Washington, DC. June. *http://eetd.lbl.gov/ea/EMS/EMS_pubs.html.*
34. Tietenberg, T. (1998). *Environmental Economics and Policy*, 2nd edition, Reading MA: Addision-Wesley Educational Publishers, Inc.
35. U.S. Department of Energy (2006). "The Benefits of Demand Response in Electricity Markets and Recommendations for Achieving Them". A Report to the U.S. Congress Pursuant to Section 1252 of the Energy Policy Act of 2005.

A
APPENDIX: A DIAGRAMMATIC WELFARE ANALYSIS OF COMPETITIVE ELECTRICITY MARKETS

In this appendix,[32] a welfare analysis of electricity markets is developed for situations where all firms can adjust their usage in response to price signals, and firms face fixed tariffs with peak and off-peak prices, using geometric diagrams.

Competitive Electricity Market with Full Capacity to Adjust to Price Signals

Consistent with much of the literature, and without loss of generality, we assume that the market for electricity is divided into two distinct periods, a peak period and an off-peak period. Further, it is a market where generators' offers to sell un-contracted capacity and energy are submitted to a last price auction. However, demand is uncertain; price is known just prior to when the quantities each generator are to serve are determined. These conditions characterize day-ahead wholesale electricity markets such as that run by the New York ISO and are consistent with the standard market design as currently proposed by FERC.

We initially assume that customers can make *full* and *costless* adjustments to demand in response to price changes according to established derived demand schedules for electricity that represent the value of the marginal product of electricity to the firm. The situation is depicted in Figure A-1. For analytical purposes below, we consider the peak and off-peak periods separately.

Off-Peak Demand

According to Figure A-1, the competitive equilibrium in the off-peak period is at point Y. Here, retail customers during off-peak periods follow a demand curve depicted as Do in the figure and buy X_3^c at price P_3^c at a total cost of $X_3^c P_3^c$. The generators supply X_3^c according to supply curve S and are paid P_3^c yielding revenue equal to $P_3^c X_3^c$. Under these conditions, welfare is measured by the sum of consumer and producer surplus:

Consumer surplus is the area under the demand curve Do and above the price line P_3^c, as indicated by the box labeled i and the triangles h and r.

Producer surplus is the area above the supply curve S and below the price line P_3^c, as indicated by (j + k + n).

Welfare is the sum of the producer and consumer surpluses, area {h + i + r} + {j + k + n}.

---

[32] This appendix is nearly identical to the analysis in an earlier paper (Boisvert and Neenan 2003) by the current authors, and it is reproduced here primarily for convenience of the reader interested in a diagrammatic treatment of the welfare analysis of competitive electricity markets. The welfare implications of markets for capacity with demand response participation in the text follow similar logic.

Peak Demand

The competitive equilibrium for the peak period if customers respond to price changes is at Z'', the intersection of the peak demand curve D_p and price P_4^c (see Figure A-1). During periods of peak demand, retail customers buy X_4^c at a price of P_4^c and a cost of $X_4^c P_4^c$. The generators supply X_4^c and are paid P_4^c, and they receive revenues of $P_4^c X_4^c$. The measure of welfare is again given by the sum of consumer and producer surplus.

Consumer Surplus is the area to left of D_p and above P_4^c, the area (a + b).

Producer Surplus is the area above S, to the left of Dp and below P_4^c, the area (h + i + r + j + k + n + s' + g).

Welfare is the area {a + b} + {h + i + r + j + k + n + s' + g}.

If this were a market for a storable commodity whose production takes place prior to knowing demand conditions, one could develop a buffer stock scheme to even out supply and demand. That is, the buffer stock agency buys when supply exceeds demand (off-peak) and sells when demand exceeds supply (supply). Under these conditions, Just, *et al*. (2008) show that society gains from such price stabilization actions if price is set at the weighted average of P_3^c and P_4^c, with the weights being the probability of each state (Just, *et al*. 2008).

Unfortunately, electricity is not storable, so the analysis of Just, *et al*. (2008) does not apply directly to these circumstances. Further, under current retail market conditions most customers can still buy electricity at fixed rates, but their suppliers face fluctuating market prices.[33] To see the value of inducing price responsiveness, we must compare the case just illustrated, where demand can fully respond to price, with the situation whereby retail customers can use any amount of electricity at fixed prices.

Competitive Wholesale Electricity Market: Retail Demand Served at Fixed Prices

To complete this part of our illustrative welfare comparison, we again look at the off-peak and peak periods separately, at least initially.

Off-Peak

We begin by examining the outcome for the off-peak period under the flat tariff, T, again assuming that demand curves are net of any wholesale margin.

In off-peak periods, the fixed tariff (T in Figure A-1) is set above the off-peak market price, because peak power is purchased at a price higher than T, and for the wholesaler to cover the cost of both peak and off-peak power purchases, T must be a weighted average of the peak and

[33] For many customers, it is not practical to adjust demand in response to price changes; the transactions costs (outage costs plus costs of administration, meters, etc.) of doing so are very high. This means that the two aggregate demand curves in Figure A-1 are the horizontal sum of many individual demand curves, most of them completely inelastic (e.g. completely vertical), or nearly so.

off-peak prices.[34] The equilibrium in this case for the customer is at point X, consuming quantity X_3^*. At point X:

Consumer Surplus = (h)

Producer Surplus = (i + j + k) (i + j go to the customer's load-serving entity (LSE); k goes to the generator)

Social Welfare = {h} + {i + j + k}

Social loss compared with the competitive market situation where customers can respond to price is: { r (foregone consumer surplus) + n (foregone producer surplus)}.

Compared with the situation described above where customers can respond to price, social welfare is reduced under the flat tariff by the areas r + n, which is called deadweight loss, while consumer surplus, area i, is transferred from customers to the LSE. Transfers do not affect the level of net social welfare, only how it is shared among consumers, generators, and retail suppliers.

To summarize, social welfare can be increased by offering to sell additional load at the lower price P_3^c. Demand and supply will continue to adjust, until the equilibrium point Y is reached. At Y:

Producer surplus increases by an amount equal to the area n

Consumer surplus increases by an amount equal to the area r, which either the supplier retains unless it lowers the price of all X_3^c to the customer, in which case the customer would realize the full benefit, and area i is transferred back to consumers.

Regardless of who retains the increase in producer and consumer surplus, Y is preferred socially to X since it represents the optimal use of resources.

Peak Period

We next examine the situation in the peak period in a similar fashion, also using Figure A-1. When customers are faced with a fixed tariff, the equilibrium point will be at point Z in Figure A-1, where the retail price is fixed at T and quantity consumed is X_4*. The flat tariff also leads to inefficiencies in the peak period because for demand greater than X_4^c, the usage price, which represents value to the firm given by points on the demand curve, is below marginal cost (e.g. the supply curve). The use of electricity whose value in production is below the cost of electricity results in deadweight loss in welfare to society represented by the combined area d + d′.

The distribution of producer and consumer surplus in the peak period case requires care to disentangle. We know that on average the price T covers the cost of the LSE's purchases of energy to serve the customers both during peak and off-peak periods. Therefore, in looking at Figure A-1, we can assume that expenditures by LSE to buy power at peak prices above T is effectively collected from the customer through off-peak sales at T which is above the supply cost, and which is then passed along to the generator. If the supply curve were indeed flat, as it

[34] As above, T is a weighted average price, where the weights are the proportion of electricity consumed in each period.

effectively is from the customer's perspective when facing a fixed price of T, consumer surplus at price T (Figure A-1) would be: a + b + g′ + f + e, and there would be no producer surplus. The wholesale suppliers and in turn generators would be paid T for each unit, and that payment would equal marginal cost.

However, implicit in the fixed tariff T (determined simultaneously with X_4*and X_3*) is a payment of X_4*[P_4*- T] (and quantity weighted) to cover the wholesaler's cost of X_4* over and above T. This amount is transferred to the generator and is equal to the combined area b + c + d + d′ + g′ + f + e. The areas b + c + f + e are consumer surplus transfers from the customer to the generator during the peak period and thus augment producer surplus above the level s′. The final result is that consumer surplus = a, and producer surplus = s′ + b +g' + c + d + d'. The generator also receives payments (economic rents) equal to the combined area d′ + d, which represents additional costs to the customer resulting from the inefficiency in pricing all usage at T rather than at the true differential prices that reflect the marginal cost of supplying electricity. From society's perspective, the additional resources needed to produce X_4* – X_4^c (e.g., consumption over and above the optimal level) would have been better allocated to other uses; thus the combined area d′ + d is lost to the detriment of society, and it is referred to as the deadweight loss.

The challenge facing the designers of electricity markets and policy makers is how to design retail programs that can reduce or eliminate altogether the size of these deadweight losses. There is perhaps no single solution to the problem, but we can highlight the important issues by illustrating the impact of a demand response program, which encourages customers to bid P_4^c to provide load reduction in the amount [X_4*– X_4^c], thereby eliminating the deadweight loss. Payments to those that accomplish this load reduction would be the combined area s″ + e + d′ (see Figure A-1). As long as this area is less than the deadweight loss of d′ + d, then social welfare is unequivocally improved. In other words, for there to be an increase in net social welfare for a DR program, (s" + e) < d; these areas are illustrated in Figure A-1.[35]

The size of these two areas is clearly an empirical question.[36] From a policy perspective, we can view this situation in two different ways. The first relates to the characteristics of supply and

---

[35] Borenstein and Holland (2002) provide an analysis of the second-best optimum if customers are to remain on flat tariffs. Their arguments are summarized here because through further analysis, one may be able to discover an algebraic relationship between these areas, although such an analysis is not done in this paper. As stated above, Borenstein and Holland (2002) shows that the quantity weighted average price, T, is the flat tariff that will cover the costs of retail electricity suppliers. However, this is not the flat tariff that provides the second-best welfare solution if retail customers stay on flat tariffs. Instead, they show that the flat rate tariff that minimizes the dead weight loss is one in which the price weights are the relative slopes of the peak and off-peak demand curves. This rate may be higher or lower than the value of T. This is an important result, but it depends on the supply curve being perfectly elastic up to system capacity, and vertical at that point. If supply elasticities are in between these extremes, the second-best fixed tariff would also likely involve the slopes of the supply curves as well, although this is not derived explicitly here. At some time it would be useful to derive this more general result, although it is not critical to the validity of their argument.
As Borenstein and Holland (2002) also point out, one difficulty with this second-best fixed tariff does not necessarily allow retail suppliers to cover their costs. However, these costs can be covered along with achieving the second-best solution under competition through a tax or subsidy that is the quantity weighted average of the new second-best flat tariff.

[36] For convenience, this Figure A-1 is drawn assuming linear supply and demand curves, but this representation may in fact distort the size of the areas being compared.

demand if firms had an incentive to respond to price and achieve the equilibrium defined by point Z'' in Figure A-1. Viewed from this perspective, it is clear that as the supply curve becomes steeper (e.g. pivoting counter clockwise around point z''), the net welfare from a demand response program increases because the area d becomes larger. Similarly, if the initial demand curve were less price responsive (made steeper by pivoting clockwise about the competitive equilibrium z'') the net welfare calculation would also move in favor of the demand response load, as the areas e and s'' would both become smaller. In summary, the potential welfare gains from demand response load programs are highest in situations where both the supply and demand curves are initially extremely price inelastic ("steeper"). These are the very circumstances that have lead to price spikes that disrupt newly formed wholesale markets.

Therefore, from a societal perspective, it makes sense to focus on exposing customers to market prices at during the peak period when they are high. This view provides a basis for understanding the size of the deadweight losses and the potential gains from implementing demand response programs. Prior to program implementation, firms would be facing a fixed tariff and consuming at point Z in Figure A-1. Thus, if we take this as a starting point, the welfare gains from a demand response program can be increased if firms: a) can be encouraged to reduce overall peak demand (e.g. resulting in a shift in D_p to the left) and/or, b) if the supply curve is sufficiently steep, firms can be encouraged to be more price responsive just during peak periods (e.g., resulting in D_p pivoting counterclockwise around point Z). The former situation calls for permanent changes in consumption patterns by introducing time-of-use pricing. The latter is more effectively accomplished by exposing customers to prices, or incentives derived there from, when such market conditions obtain.

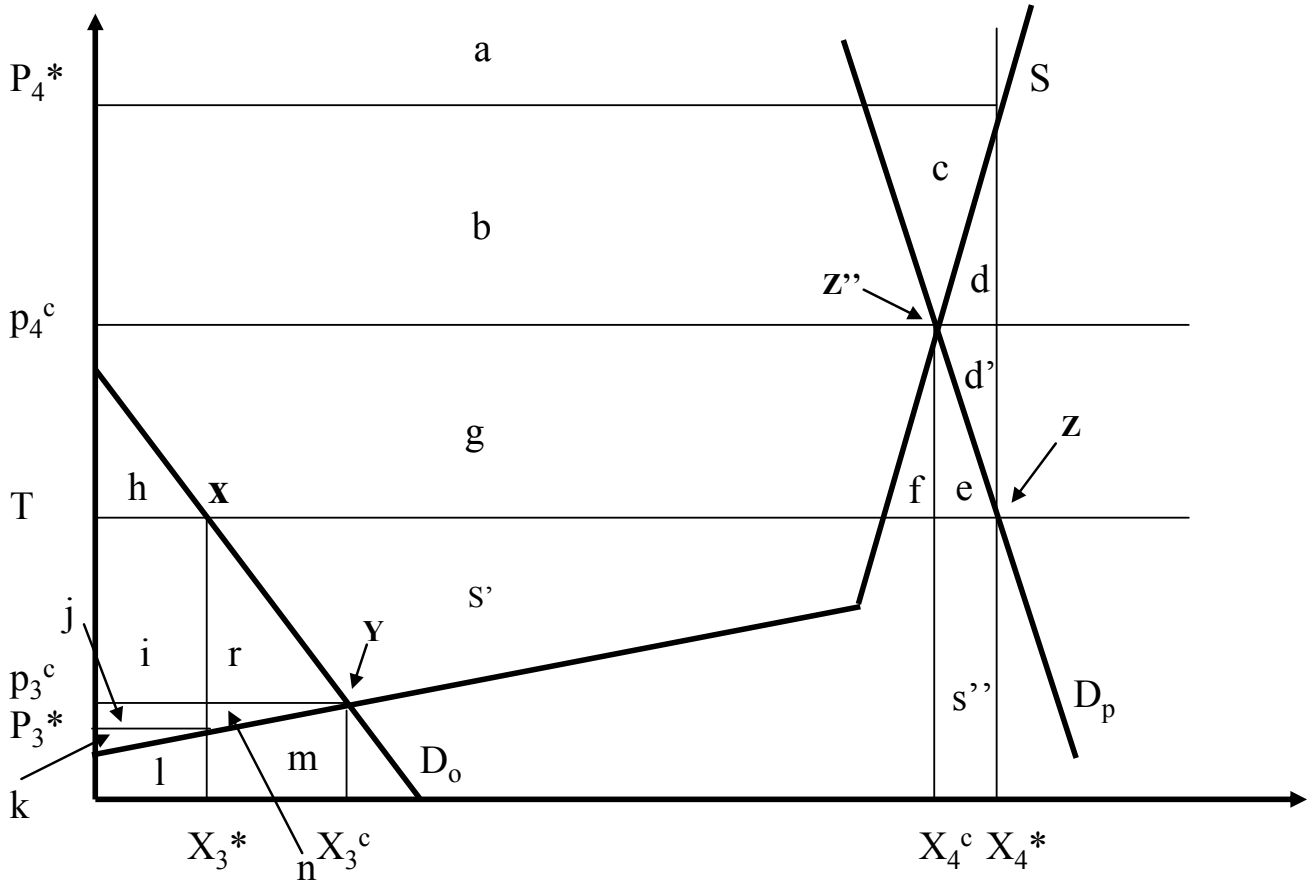


Figure A-1

B

APPENDIX: MARKET EQUILIBRIUM WITH OTHER STYLIZED EXAMPLES OF DEMAND RESPONSE SUPPLY CURVES FOR CAPACITY

The purpose of this appendix is to illustrate that the size of the social deadweight losses avoided through demand response programs that allow electricity customers to supply capacity by dropping load depends critically on the nature of their supply curves. There are two examples reported in this appendix. One example is illustrated in Figures 10-1 through 10-4. The second example is illustrated in Figures 10-5 through 10-8.

The Two Cases

These two examples differ from the one is Section 5 of the text primarily by the size of the first increment of capacity offered in the first step of the supply function. In the first example (Figure B-1), the size of the first increment offered at a price of P_{0DR} is considerably smaller than the one in Figure 5-4 of the text, while in the second (Figure B-5), the initial increment offered at a price of P_{0DR} is significantly larger than the one depicted in Figure 5-4 of the text. For comparison purposes, the generator supply curves, and the market demand curves for capacity are assumed to be the same as in the discussion in the text.

To make these comparisons, one can begin with the first example. This demand curve is depicted in Figure B-1 is superimposed on the market for capacity in Figure B-2. Then, following logic similar to that in Section 5 above, it is clear from Figure B-2 that once demand response participants are allowed to offer capacity into the market, the new equilibrium is established at the point labeled E^{14A}_{G+DR}. Compared with the equilibrium that would have obtained without this additional supply of capacity, the market clearing price has dropped from P_G to P^{14A}_{G+DR}, and the quantity of capacity supplied by generators has dropped from $Q_G = Q_M$ to Q^{14A}_{G}, while demand response participants now supply an amount equal to Q^{14A}_{DR}. Because the initial increment offered is so small, the market clearing quantity of capacity offered by the demand response customers includes a portion of the amount offered in the second step (Figure B-2). This is why the drop in the equilibrium price is slightly smaller than in the case examined in Section 5. Given the small size of the initial increment, compared with that in Section 5, it is not surprising that the size of the social deadweight avoided by allowing demand response participants to offer capacity is smaller as well. One only needs compare the green shaded areas in Figure 5-5 and Figure B-2 to establish this fact.[37]

To continue with the comparison, the demand curve for the second example is depicted in Figure B-5, and it is superimposed on the market for capacity in Figure B-6. Again following logic similar to that in Section 5 above, it is clear from Figure B-6 that once demand response

[37] There is little to be gained by examining Figures 10-3 and 10-4 in detail. However, through an analysis similar to that for Figures 5-6 and 5-7 in the text, these two figures also illustrate how the supply of capacity through demand response participants can help alleviate the problems due to unforeseen events that would lead to outward shifts in demand for capacity or reductions in the supply available from generators.

participants are allowed to offer capacity into the market, the new equilibrium is established at the point labeled E^{14B}_{G+DR}. Compared with the equilibrium that would have obtained without this additional supply of capacity, the market clearing price has dropped from P_G to P^{14B}_{G+DR}, and the quantity of capacity supplied by generators has dropped from Q_G to Q^{14B}_{G}, while demand response participants now supply an amount equal to Q^{14B}_{DR}. Because the initial increment offered is so large, the market clearing quantity of capacity offered by the demand response customers need not include a portion of the amount offered in the second step (Figure B-6). This is why the drop in the equilibrium price is somewhat larger that than in the case examined in Section 5, and considerably larger than in example 1 in this appendix. Given the large size of the initial increment, compared with that in Section 5, it is not surprising that the size of the social deadweight avoided by allowing demand response participants to offer capacity is larger as well. One only needs compare the green shaded areas in Figure 5-5 and Figure B-6 to establish this fact.[38]

A Summary

As stated above, the purpose of this appendix is to illustrate that the size of the social deadweight losses avoided through demand response programs that allow electricity customers to supply capacity by dropping load depends critically on the nature of their supply curves. To some extent, these two cases could be viewed as extremes. The first is one in which the supply of capacity at low prices from demand response participants is modest. In contrast, the second is one in which the supply of capacity at low prices from demand response participants is rather large. In this sense, the case discussed in the text is one that lies in between these extremes. The lessons to be drawn from these comparisons are that regardless of the shape of these supply curves, the qualitative effects of this additional source of supply of capacity are the same. A major consequence is to reduce the equilibrium price of capacity. To the extent that this is so, the revenues generated by generators in this market will be diminished, and this may work to slow the rate of future investment in new generation capacity. However, this is as it should be. It demonstrates that since customers are willing to supply some capacity through load reduction, they place a lower value on investment in future generation capacity, and thus, any reduction in the rate of future investment in new generation is an efficient market outcome.

There is one other point to emphasize. The one example demand response supply curve for capacity discussed in Section 5 and the two outlined here represent a range in the behavior that could be exhibited by a heterogeneous mix of customers in any electricity market. And, in the aggregate, the market supply curve for capacity by demand response customers would be the horizontal summation of the supplies by the diverse set of customers. In the aggregate, this supply curve for capacity is likely to appear more as a continuous function with a positive slope when measured from right to left on a figure depicting a market for capacity. Alternatively, because of the one-to-one joint production of capacity and load reduction, movements along this supply curve from left to right also trace out a continuous downward sloping demand curve for capacity. It is this type of aggregate supply/demand curve for capacity that is used in Section 6 to examine the effects of demand response supplied capacity in an ISO/RTO run capacity market.

---

[38] There is also little to be gained by examining Figures 10-7 and 10-8 in detail. However, through an analysis similar to that for Figures 5-6 and 5-7 in the text, these two figures also illustrate how the supply of capacity through DR participants can help alleviate the problems due to unforeseen events that would lead to outward shifts in demand for capacity or reductions in the supply available from generators.

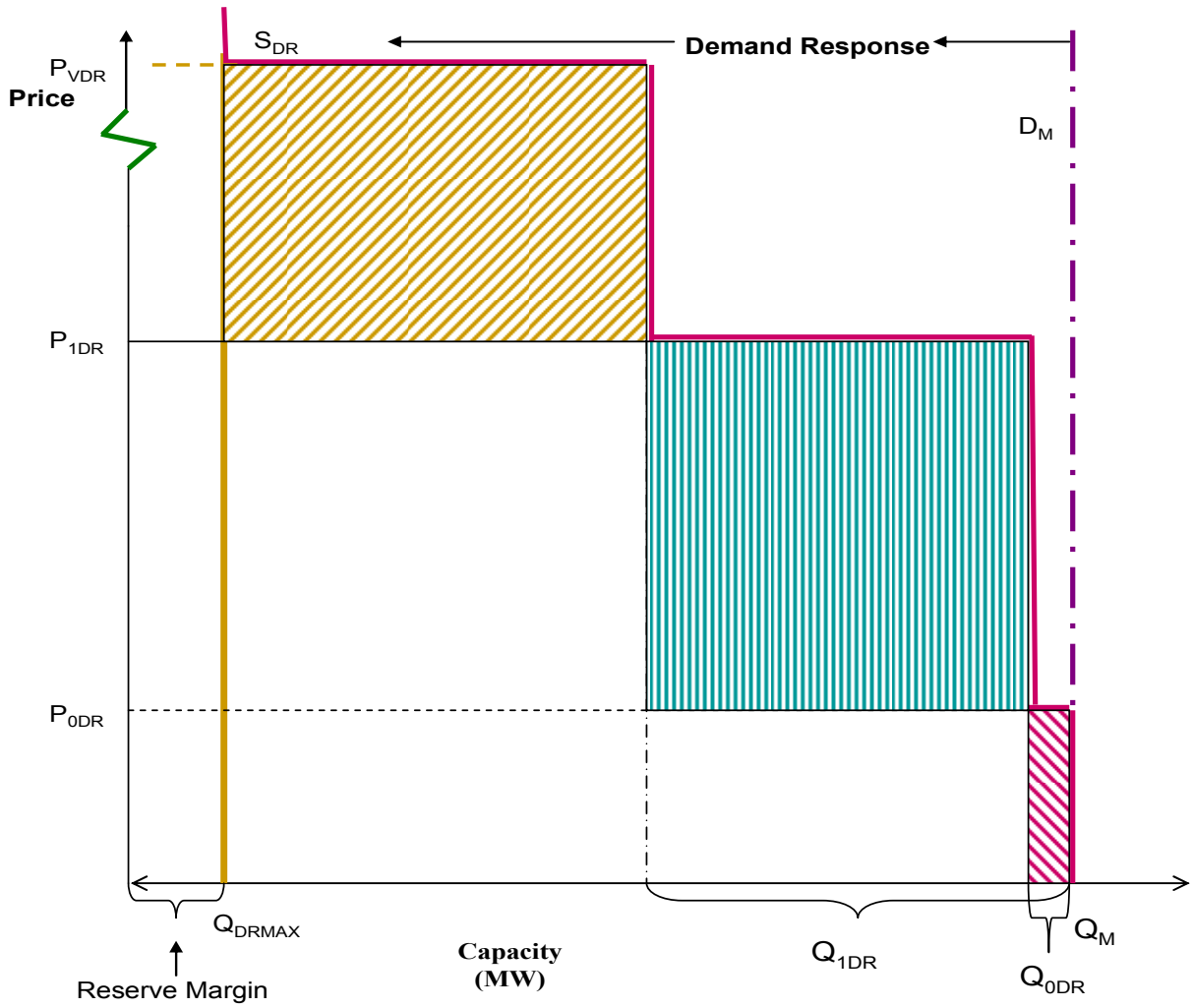

S_{DR}
Demand Response
P_{VDR}
Price
D_M
P_{1DR}
P_{0DR}
Q_{DRMAX}
Reserve Margin
Capacity
(MW)
Q_{1DR}
Q_M
Q_{0DR}

Figure B-1

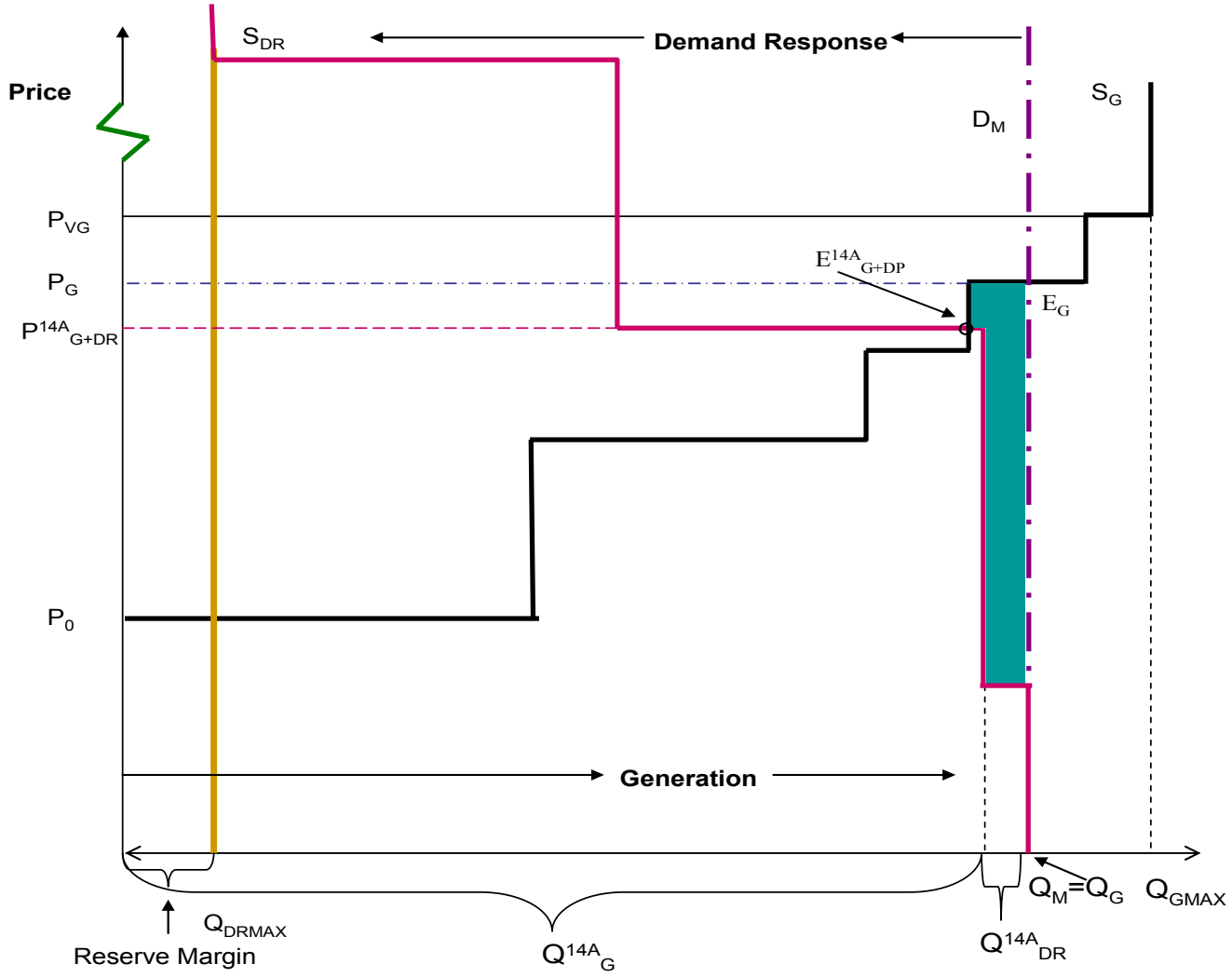

Price
S_{DR}
Demand Response
S_G
D_M
P_{VG}
P_G
P^{14A}_{G+DR}
E^{14A}_{G+DP}
E_G
P_0
Generation
Q_{DRMAX}
Reserve Margin
Q^{14A}_G
Q^{14A}_{DR}
$Q_M=Q_G$
Q_{GMAX}

Figure B-2

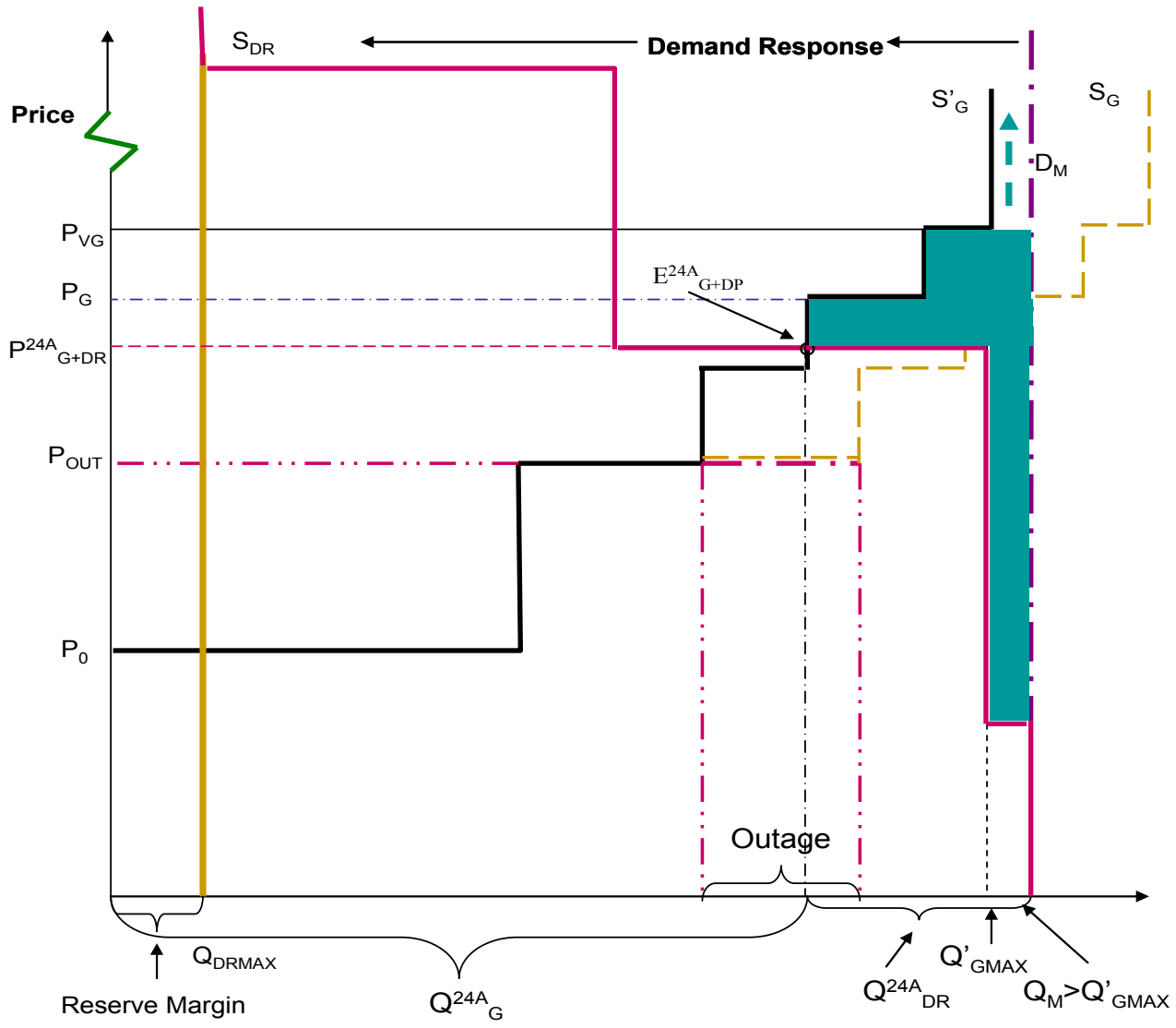

S_DR
Demand Response
Price
S'_G
S_G
D_M
P_VG
P_G
E^24A_G+DP
P^24A_G+DR
P_OUT
P_0
Outage
Q_DRMAX
Reserve Margin
Q^24A_G
Q^24A_DR
Q'_GMAX
Q_M>Q'_GMAX

Figure B-3

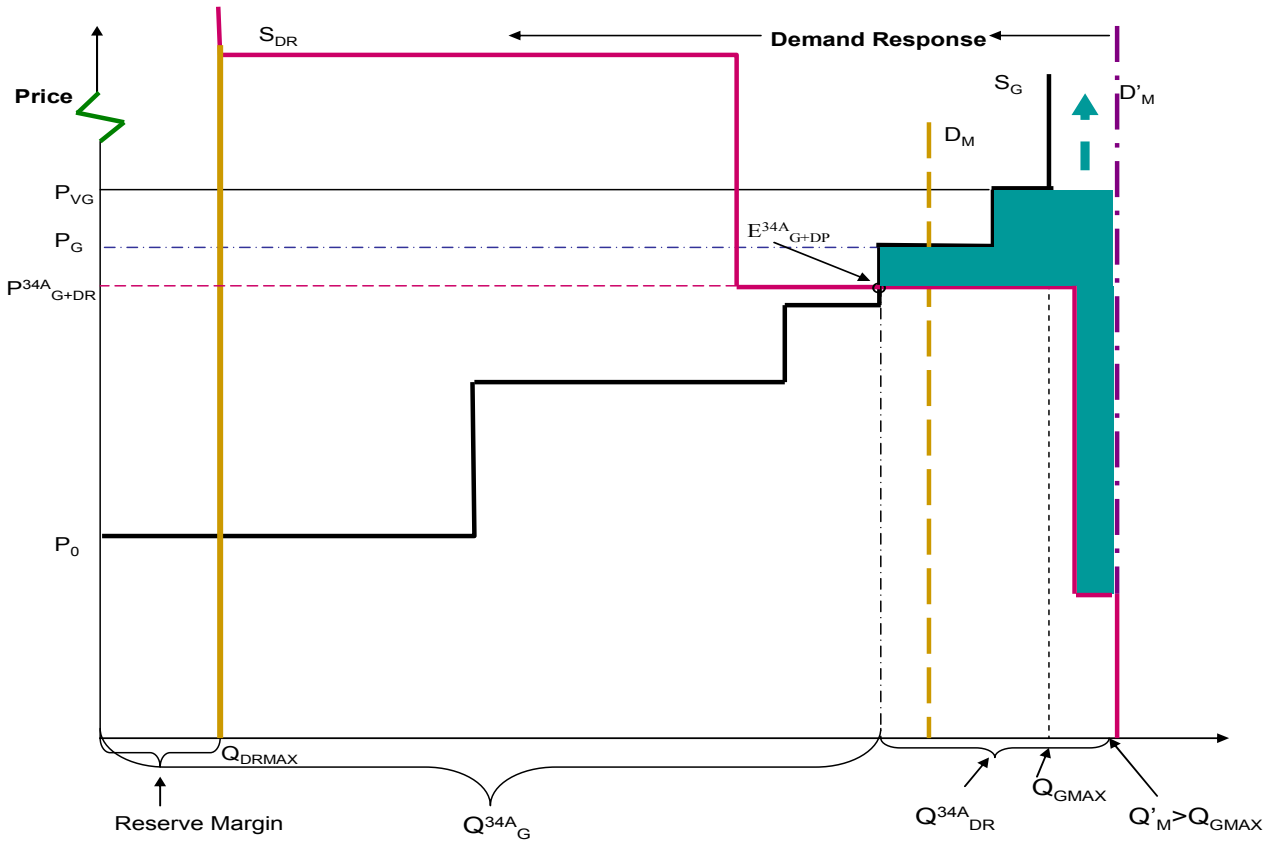

Price
S_{DR}
Demand Response
S_G
D'_M
D_M
P_{VG}
P_G
E^{34A}_{G+DP}
P^{34A}_{G+DR}
P_0
Q_{DRMAX}
Reserve Margin
Q^{34A}_G
Q^{34A}_{DR}
Q_{GMAX}
$Q'_M > Q_{GMAX}$

Figure B-4

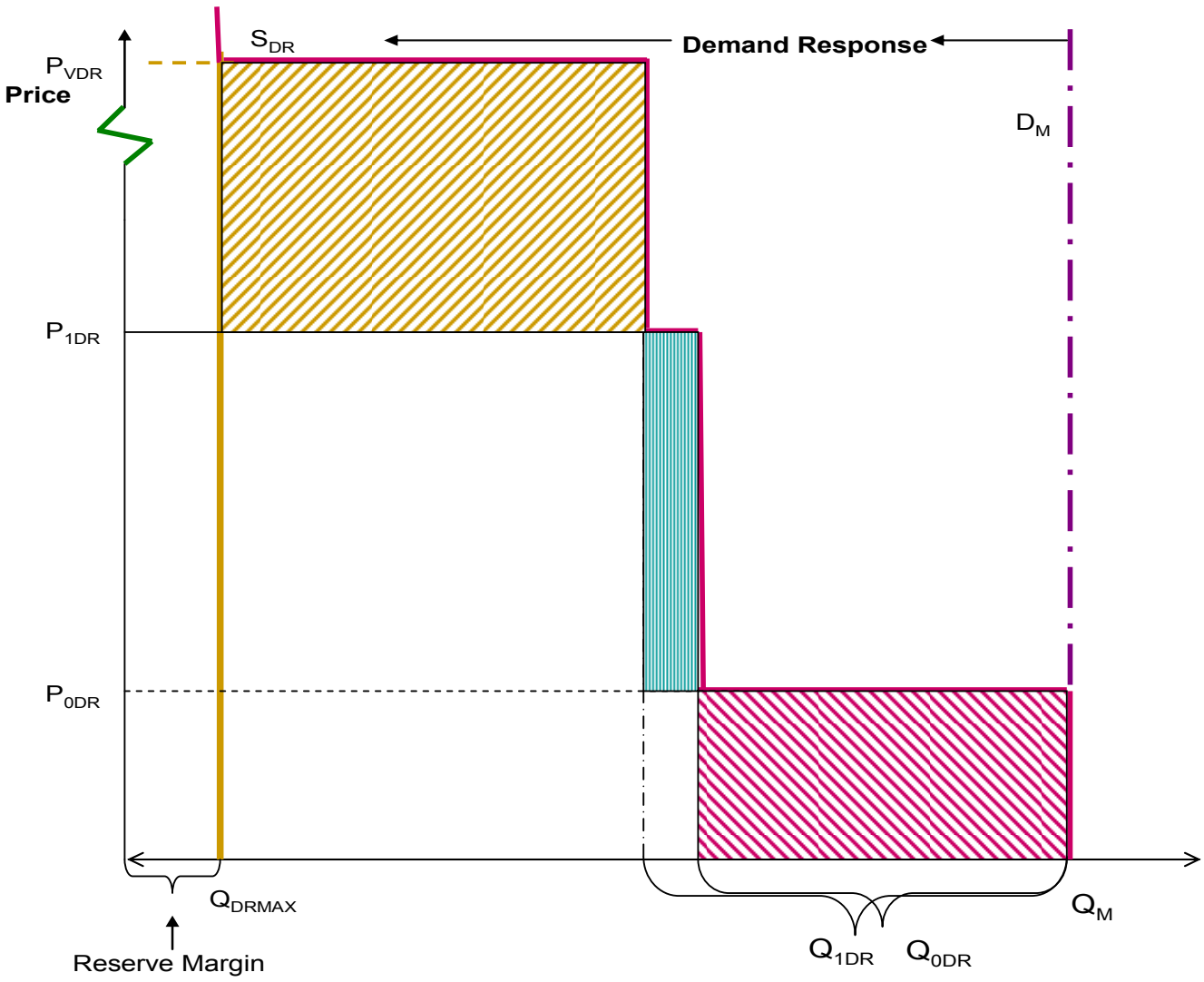

S_{DR}
Demand Response
P_{VDR}
Price
D_M
P_{1DR}
P_{0DR}
Q_{DRMAX}
Reserve Margin
Q_{1DR}
Q_{0DR}
Q_M

Figure B-5

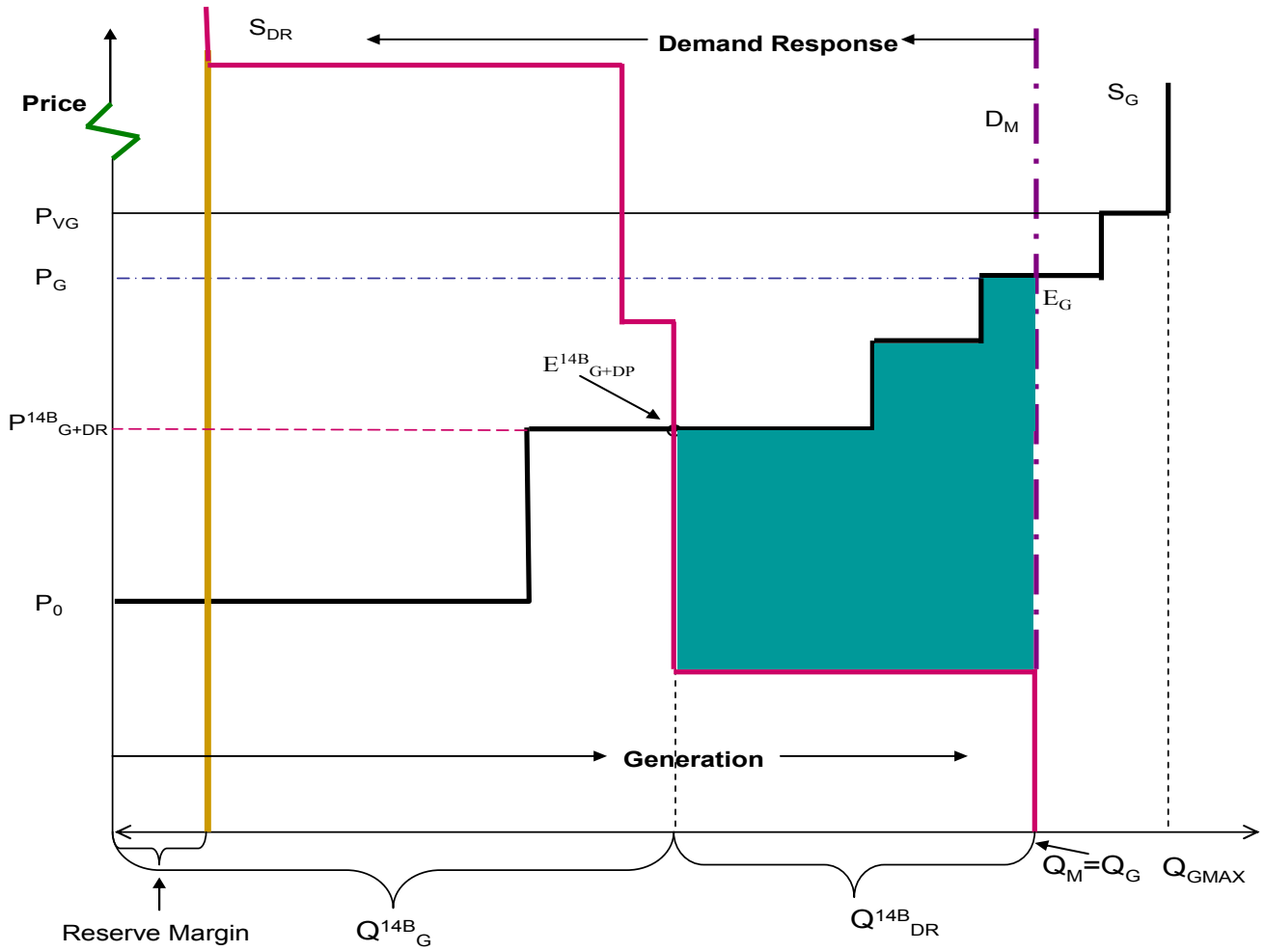

S_DR
Demand Response
Price
D_M
S_G
P_VG
P_G
E_G
E^14B_G+DP
P^14B_G+DR
P_0
Generation
Q_M=Q_G
Q_GMAX
Reserve Margin
Q^14B_G
Q^14B_DR

Figure B-6

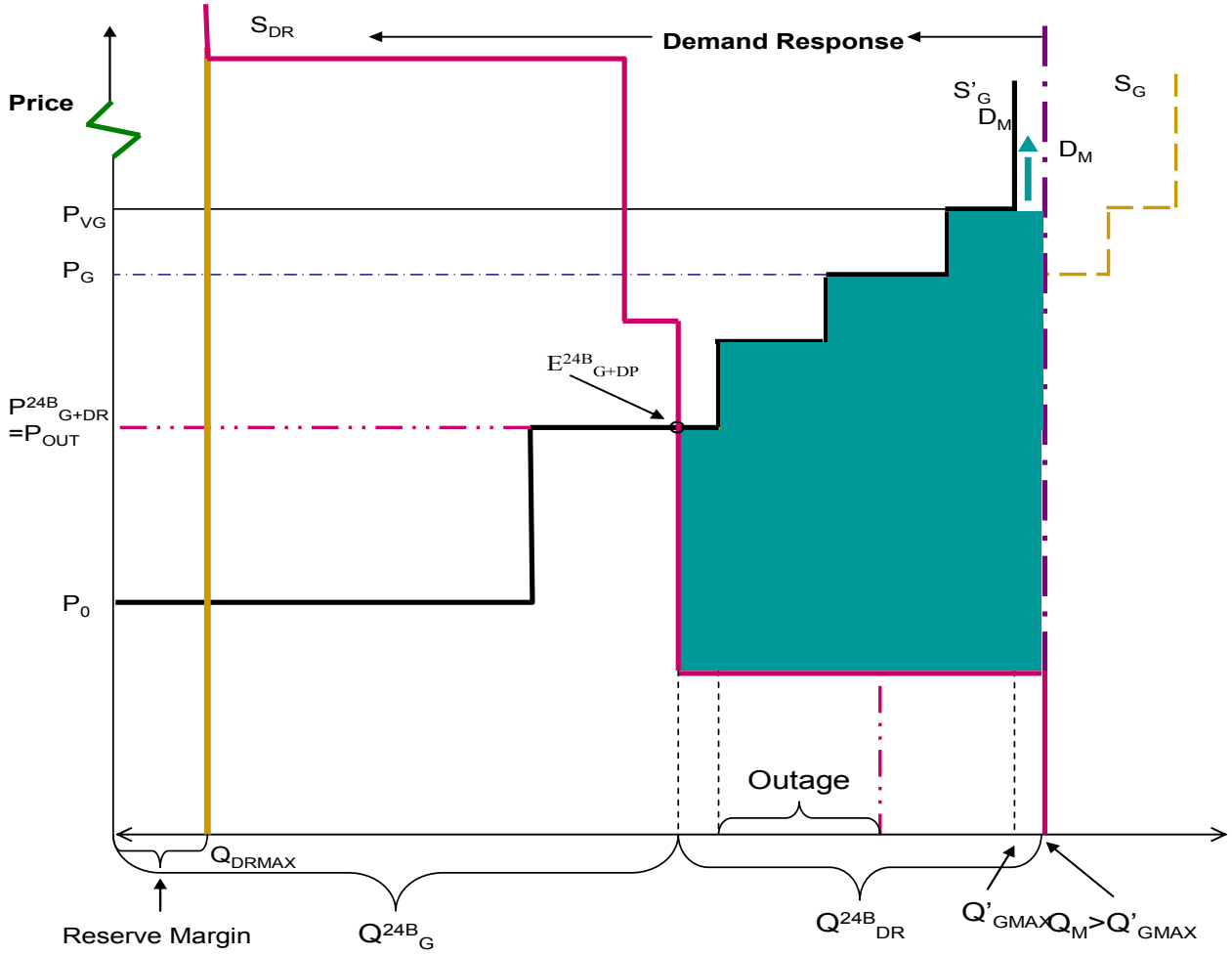

S_DR
Demand Response
Price
S'_G
D_M
S_G
D_M
P_VG
P_G
E^24B_G+DP
P^24B_G+DR
=P_OUT
P_0
Outage
Q_DRMAX
Reserve Margin
Q^24B_G
Q^24B_DR
Q'_GMAX
Q_M>Q'_GMAX

Figure B-7

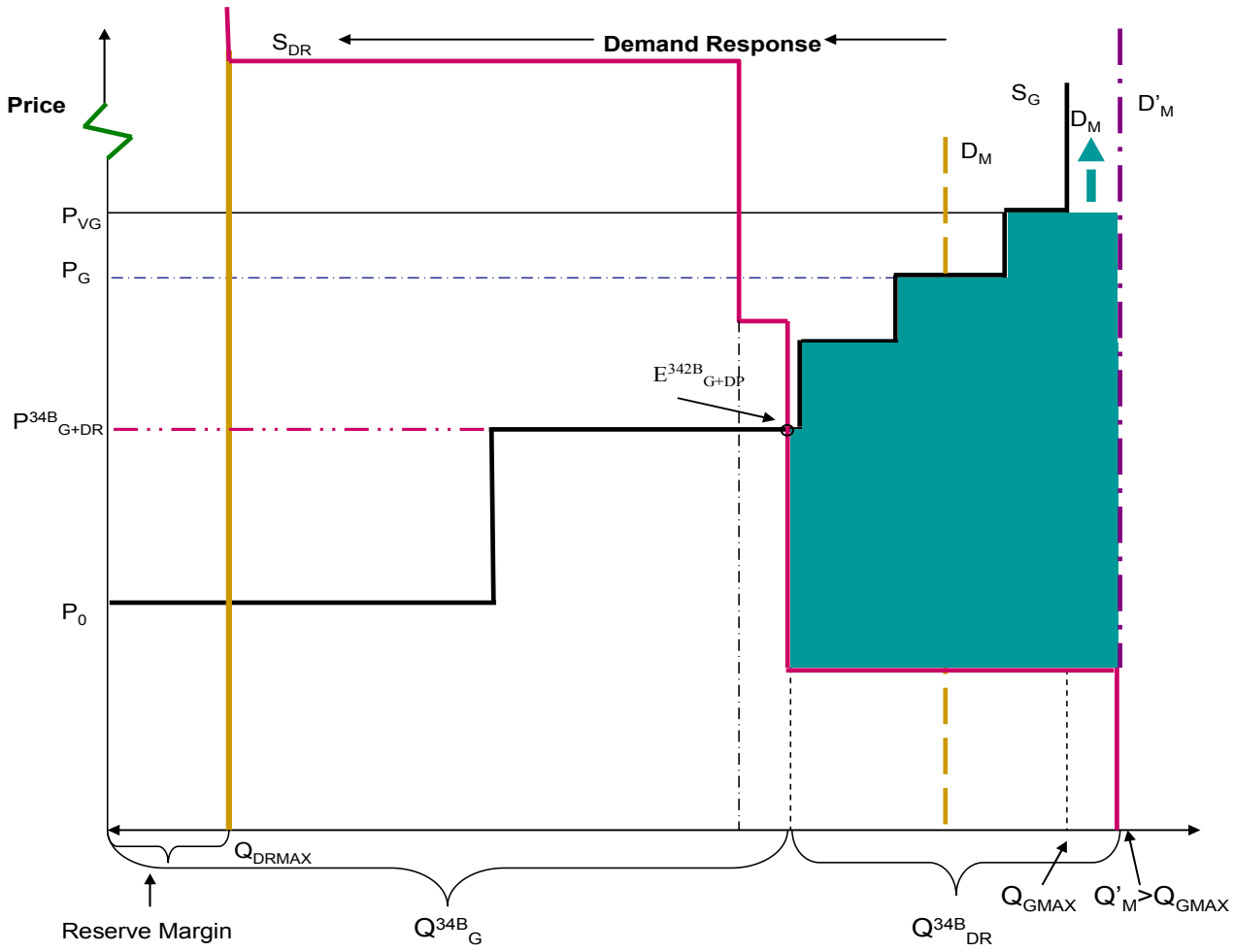

Price
S_{DR}
Demand Response
S_G
D_M
D'_M
D_M
P_{VG}
P_G
E^{342B}_{G+DP}
P^{34B}_{G+DR}
P_0
Q_{DRMAX}
Reserve Margin
Q^{34B}_G
Q^{34B}_{DR}
Q_{GMAX}
$Q'_M > Q_{GMAX}$

Figure B-8

Export Control Restrictions

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

The Electric Power Research Institute Inc., (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, health, safety and the environment. EPRI also provides technology, policy and economic analyses to drive long-range research and development planning, and supports research in emerging technologies. EPRI's members represent more than 90 percent of the electricity generated and delivered in the United States, and international participation extends to 40 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together…Shaping the Future of Electricity

1017876

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com