Prepared for:

Alberta Utilities Commission
400, 425 First Street S.W.
Calgary, AB T2P 3L8

The Economic Foundations of Capacity Markets

Prepared by:

Charles River Associates
80 Bloor St. West
Toronto, Ontario, M5S 2V1
www.crai.com/energy

Date: June 2, 2017
CRA D23420

Acknowledgements

This report was authored by Brian Rivard, Jordan Kwok and Neve Sterns.

Disclaimer

The views expressed herein are the views and opinions of the authors, and do not reflect or represent the views of Charles River Associates or any of the organizations with which the authors are affiliated. Any opinion expressed herein shall not amount to any form of guarantee that the authors or Charles River Associates has determined or predicted future events or circumstances, and no such reliance may be inferred or implied. The authors and Charles River Associates accept no duty of care or liability of any kind whatsoever to any party, and no responsibility for damages, if any, suffered by any party as a result of decisions made, or not made, or actions taken, or not taken, based on this report.

Table of Contents

List of Acronyms

AESO	Alberta Electric System Operator
AUC	Alberta Utilities Commission
CAIDI	Customer Average Interruption Duration Index
CAIFI	Customer Average Interruption Frequency Index
CAISO	California Independent System Operator
CLP	Climate Leadership Plan
CONE	Cost of New Entry
DCA	Descending Clock Auction
DR	Demand Response
EFORd	Equivalent Forced Outage Rate
ERCOT	Electric Reliability Council of Texas
EUE	Expected Unserved Energy
FERC	Federal Energy Regulatory Commission
IRC	ISO/RTO Council
ISO-NE	New England ISO
LOLE	Loss of Load Expectation
LOLEV	Loss of Load Event
LOLH	Hourly Loss of Load Expectation
LOLP	Loss of Load Probability
LSE	Load Serving Entity
MISO	Midcontinent ISO
MOPR	Minimum Offer Pricing Rule
NYISO	New York ISO
ORDC	Operating Reserve Demand Curve
PJM	PJM Interconnection
RO	Reliability Option
RTO	Regional Transmission Organization
SAIDI	System Average Interruption Duration Index
SAIFI	System Average Interruption Frequency Index
SBA	Sealed Bid Auction
UCAP	Unforced Capacity
VOLL	Value of Lost Load

Executive Summary

This report, prepared for the Alberta Utilities Commission, examines the economic and theoretical foundations of capacity markets. There is empirical evidence that “energy-only” competitive wholesale markets may not provide sufficient incentive for private firms to invest in generation to the levels required to maintain established industry resource adequacy requirements. This issue has been called the “resource adequacy" problem or the “missing money” problem. The missing money problem is driven by a number of factors that lead to the broader market failure, including the pursuit of physical rather than economic reliability standards, wholesale market imperfections, regulatory constraints that drive spot market outcomes, and system operator interventions. Academics and practitioners have proposed and implemented various solutions to address the resource adequacy and missing money problems. A capacity market is one of the most often implemented.

This report discusses the various causes of the resource adequacy and missing money problems, and the various approaches proposed to address the issues. The report also considers the basic capacity market design principles and features needed to achieve efficient outcomes, thereby realizing the benefits associated with competition.

1. Introduction

The Alberta Electric System Operator (AESO) currently operates an "energy-only" market in which the energy price provides the underlying driver for new investment.[1] The energy-only market has functioned well to date, attracting needed new investment, maintaining a reliable supply of electricity, and providing competitively priced electricity.[2]

In November 2015, the Government of Alberta introduced the Climate Leadership Plan (CLP) to reduce carbon emissions.[3] The CLP will significantly change the future supply mix in Alberta. All coal generation will be shut down by 2030. Replacement of the coal generation assets involves a planned increase of 5,000 MW of new renewable energy generation assets. Additional investment in the form of natural gas generation or other dispatchable resources will also be required.

The AESO recently conducted an analysis of the ability of the current energy-only market to meet the CLP objectives, examining whether a reliable supply of new generation would emerge through private investment stimulated by competitive prices.[4] Based on this analysis, the AESO recommended that the province move away from an energy-only market structure toward a structure that includes both an energy market and a capacity market. The AESO anticipates that the earliest date that capacity could be procured by using an auction process would likely be 2024, and a bridging mechanism will likely be required to ensure supply adequacy between the years 2021 and 2024.[5]

This report, prepared for the Alberta Utilities Commission (AUC), examines the economic and theoretical foundations of capacity markets. The report explores three related topics:

- The economic and reliability rationales for why capacity mechanisms exist within the context of competitive electricity markets;
- The different forms of capacity mechanisms along with their related pros and cons; and
- The basic capacity market design principles and features needed to achieve efficient outcomes and, thereby, realize the benefits of competition.

---

1 More precisely, the AESO operates a real-time energy market, a day-ahead operating reserve market, and procures several other ancillary services through contracts. In the literature, it is implicitly assumed that an "energy-only market" includes associated ancillary service markets and procurement processes. This is discussed in more detail in Section 2.2.

2 See https://www.aeso.ca/assets/Uploads/Albertas-Wholesale-Electricity-Market-Transition.pdf at p. 5.

3 See https://www.alberta.ca/climate-leadership-plan.aspx.

4 See https://www.aeso.ca/assets/Uploads/Albertas-Wholesale-Electricity-Market-Transition.pdf.

5 *Id.* at p. 4. The AESO also indicates that in addition to the design of the capacity market itself, the effect of a capacity market structure on the existing energy market will need to be addressed.

This report addresses these topics through a review of the academic literature and studies used to inform the adoption and development of capacity mechanisms in various jurisdictions. Each topic is discussed below.

2. Economic and Reliability Rationales for Capacity Markets

The key objective of competitive electricity market restructuring in the 1980s and 1990s was to promote economic efficiency. The lower costs associated with efficient market outcomes were expected to be realized in the short term, through improved dispatch, in the medium term, through improvements in the operating performance and labour productivity of the existing generating facilities, and most importantly, over the long term, through cost savings associated with new generation investment. Underlying these lower costs and higher productivity was the economic theory that a single, short-term marginal energy price was all that was required to stimulate efficient operation and move private investment towards an optimal mix of generation assets.[6]

However, as jurisdictions across the world gained experience with single-price, or energy-only, competitive markets, many called into question the ability of these markets to provide sufficient incentive for private firms to invest in generation to the levels required to ensure maintenance of traditional resource adequacy standards.[7] The perceived inability of the energy-only market to incentivize the quantity of investments required to maintain traditional resource adequacy standards is termed the "resource adequacy" problem.[8] Several jurisdictions, particularly those in the United States, saw capacity markets as a solution to the resource adequacy problem. Capacity markets provide generators with additional revenue based on a price for "capacity" to encourage private investment at levels required to ensure resource adequacy.

Capacity, as a separate product, is not a product for which a market would naturally arise; its existence is a consequence of the establishment of reliability criteria, specifically the standards set to ensure resource adequacy.[9] This section begins with a discussion of traditional resource adequacy standards and the trade-offs between reliability and economic efficiency associated with their use. This is followed by a review of

6 See Paul Joskow, "Restructuring, Competition and Regulatory Reform in the U.S. Electricity Sector," *Journal of Economic Perspectives*, Vol. 11, No. 3, Summer 1997, pp. 119–138.

7 For a discussion of the observed investment problem, see Paul Joskow, "Markets for Power in the United States: An Interim Assessment," *The Energy Journal*, 27(1), 2006.

8 See Peter Cramton and Steven Stoft, "The Convergence of Market Designs for Adequate Generating Capacity," white paper for the California Electricity Oversight Board, March 2006.

9 See Paul Joskow, "Symposium on 'Capacity Markets,'" *Economics of Energy & Environmental Policy*," Vol. 2, Issue 2, September 2013. This is also true of ancillary service products such as operating reserve, regulation, and voltage support, the standards for which serve to ensure system security.

the various economic explanations for why energy-only markets will often fail to incentivize sufficient private investment to meet the reliability standard and, hence, require the addition of an administrative capacity mechanism, or market, to supplement the primary market in energy.

2.1. Resource Adequacy Standards

A physical resource adequacy standard is one that is focused on system outcomes rather than economic objectives. Discussion will focus on the "one day in 10 years" (alternatively, "one event in 10 years" or "1-in-10)" loss of load expectation (LOLE) criterion, as this is an approach that has been widely adopted, particularly in North America. This is followed by a description of how resource adequacy standards are translated into system planning and reserve margins, and the trade-offs between reliability and economic efficiency associated with determining reserve margin objectives based on a physical resource adequacy standard.

2.1.1. Physical Resource Adequacy Standards

The reliability of the electric power system is generally described in terms of adequacy and security. Adequacy refers to the ability of the system to provide sufficient supply to serve firm demand in aggregate. Security pertains to the ability of the system to withstand disturbances, be they natural or man-made. Capacity mechanisms address the issue of adequacy and the associated resource adequacy standards employed by system operators. In understanding what a resource adequacy standard is, it is instructive to first understand what it is *not*:

- Resource adequacy standards are **not a measure of transmission reliability**. While resource adequacy analysis takes account of transmission constraints that may limit deliverability over the transmission system, the adequacy of the transmission system is generally defined by system security requirements and is ensured via other processes more closely associated with system stability. Accordingly, failures of the transmission system do not count against resource adequacy metrics. Transmission events beyond design contingencies typically lead to uncontrolled and potentially widespread outages; supply inadequacy manifests in controlled "brownouts" where, in the absence of available incremental supply, system operators shed firm load selectively to balance supply and demand.
- Resource adequacy standards are **not a measure of distribution reliability**. Most electric service disruptions are caused by disruptions in the distribution system when storms or other incidents cause component failures or downed low-voltage lines. Distribution events generally lead to uncontrolled,

localized outages[10] and, like transmission events, are not considered as failure to procure adequate supply resources.

- Resource adequacy is generally **measured in aggregate, not as observed by individual loads**. Lost load, or unserved energy, is assessed across the entire electric system, as are counts of loss of load events. Only a small subset of customers may be affected by each outage, even though that outage will count, on an absolute basis, against system performance. Accordingly, any individual customer is likely to observe resource adequacy at levels better than the system as a whole.[11]
- Physical resource adequacy standards are generally **not determined using economic analysis**. Though certain costs can be implied from a physical criterion, in many cases such approaches to resource adequacy do not reflect any explicit cost-benefit analysis or value of lost load (VOLL) calculation, nor do they consider least cost operation of the power system.[12]

Turning to what resource adequacy standards *are*: they are generally an expression of the acceptable frequency or duration of interruptions of power to firm load caused by insufficiency of supply resources. Resource adequacy studies are performed on a forward-looking basis to determine how much supply is expected to be necessary to realize the target metric. Resource adequacy studies are generally probabilistic in nature, and consider a number of factors, including quantity and shape of load, availability of demand response (DR) resources, transmission constraints and import availability, characteristics of interconnected renewable energy resources, and generator size, type, and outage rates. These analyses are then used to define overall resource requirements.

2.1.2. One-in-Ten

The most common North American resource adequacy standard is the 1-in-10 LOLE criterion; this approach offers the richest history and background. References to the 1-in-10 criterion can be found in utility literature

10 Importantly, the public perception of distribution outages is often more sympathetic, and consumers understand that certain acts of nature have hard-to-avoid consequences. However, outages due to inadequate procurement of resources are more likely to be perceived as mismanagement on the part of the utility or regulator.

11 See James F. Wilson, "One Day in Ten Years? Resource Adequacy for the Smart Grid," *Wilson Energy Economics*, November 2009, p.13.

12 In Great Brittan and Ireland, the resource adequacy target is informed by certain economic analyses related to VOLL and the cost of new entry (CONE) of a benchmark generation technology, in addition to political and other institutional considerations. See the discussion in "A Case Study in Capacity Market Design and Considerations for Alberta," a CRA report prepared for the AESO at p. 18, available at https://www.aeso.ca/assets/Uploads/CRA-AESO-Capacity-Market-Design-Report-03302017-P1.pdf.

dating back to the 1940s.[13] Original literature does not provide a justification for the criterion, other than the unsupported assertion that it was the level to which customers were already accustomed. Also, given the lack of modern computing capabilities, it was likely a standard that provided planners and operators with sufficient excess capacity to be confident that reliable operation of the system would remain possible. At the time, energy demand was growing at a relatively steady state and generation additions primarily included slow-to-build, long-lived assets. Under these circumstances, the cost of not constructing over time was substantial and building capacity well in advance was sensible, as it would all eventually be put to use. Furthermore, providing a high level of resource adequacy was (as it continues to be) generally a mandate of resource planners, who were provided with the means to place a greater emphasis on reliability, for which they were responsible, than on cost, for which they were not, as these costs could be passed through to customers. As a result, some would argue that, although the 1-in-10 criterion was accepted in principle, its application has been executed in such a way that curtailments of load were never actually expected to occur.[14]

More recently, some experts have started to question the continued relevance of 1-in-10 and LOLE as an acceptable criterion for resource adequacy. In addition to potential conflicts with wholesale market design and questionable economic implications, both of which will be discussed in more detail in the sections that follow, many of the historical conditions that facilitated this approach to resource adequacy are no longer relevant. Load growth has slowed significantly as a result of economic, technological and policy changes (e.g., conservation programs, promotion of DR and energy efficiency standards). Most incremental capacity additions now come more from short lead-time resources like DR, generation upgrades and gas turbines, rather than from new construction of long lead-time resources like coal and nuclear power plants.[15] Finally, developments in smart grid technology, DR programs and other forms of customer engagement now promise a more involved demand side that is better able to express its willingness to pay for reliability. Thus, while 1-in-10 could once be justified in spite of its lack of economic underpinnings, many of the rationales that supported its continued use no longer hold.

13 NERC suggests the first time such a standard appears in the literature is in a paper presented by Giuseppe Calabrese at the IAEE Midwest Generation Meeting in Chicago in 1947. Calabrese's work suggested that power systems with an LOLE index of 1-in-10 or below were easier to operate at a high level of performance than systems with indexes greater than 1-in-10. See NERC Integration of Variable Generation Task Force, "Methods to Model and Calculate Capacity Contributions of Variable Generation for Resource Adequacy Planning," March 2011, p. 14. *See also*, G. Calabrese, "Determination of Reserve Capacity by the Probability Method," Transactions of the AIEE, Vol. 69, No. 2, pp. 1681–1689, January 1950.

14 See supra note 11, at 7.

15 For example, in PJM's first six base residual auctions for their capacity markets, 82% of cleared resources were considered to have short lead times. See supra note 11 at 20.

2.1.3. Alternative Measures of Resource Adequacy

One-in-Ten has been interpreted differently by various planners and regulators. As such, discussions of 1-in-10 may refer to different approaches with different advantages and disadvantages; each approach captures — or fails to capture — the relevant parameters of shortfall events: frequency, duration and magnitude. Moreover, there are other criteria that may be considered as resource adequacy standards, each with its own strengths and weaknesses. Approaches include:

- A **daily LOLE** criterion specifies the expected number of days over a specified time period during which there is a capacity shortfall. A 1-in-10 daily LOLE could also be thought of as one instance in 10 years, or an average of 0.1 expected shortages each year. It should be noted that, when interpreted in this fashion, LOLE does not account for the duration or magnitude of a shortfall. The daily LOLE approach may be referred to as a loss of load event (LOLEV) standard, and is the most common interpretation.
- An **hourly LOLE (LOLH)** metric counts the expected number of hours during a particular period, rather than the number of times, during which load is expected to exceed resource capabilities. This interpretation of 1-in-10 would allow for 24 cumulative hours of hourly LOLE every 10 years.[16] This metric requires more data, but provides a more precise indication of the expected level of reliability as it accounts for both frequency and duration. As is the case with the daily LOLE criterion, the hourly LOLE criterion does not indicate anything about the magnitude of service interruptions. An hourly LOLE metric can be converted to a loss of load probability (LOLP), an equivalent metric, but in terms of the probability that supply will be inadequate to serve demand over a particular time period.
- **Expected Unserved Energy** (EUE) is a resource adequacy metric that measures the expected quantity of energy demand, measured in MWh, that will not be served over a specified period. This less common metric has been applied more often in systems with large amounts of hydropower capacity. EUE has the advantage of considering the variability of load and resources during all periods. Moreover, because of the units in which it is measured, EUE enables a more direct comparison between any unserved load and economic valuations, which can then be used to assess the economic effect of lost load. While EUE accounts for the magnitude and duration of resource shortfalls, it does not provide information on their frequency.[17]

16 A study commissioned by the Federal Energy Regulatory Commission and executed by The Brattle Group suggests that the difference between one day in 10 years and one event in 10 years can amount to more than a 5% difference in the recommended installed reserve margin. See Johannes P. Pfeifenberger, et al., "Resource Adequacy Requirements: Reliability and Economic Implications," The Brattle Group, Astrape Consulting, September 2013, p.iii.

17 See "Energy Division Proceeding Status Update and PRM Modeling Manual," CPUC proceeding R.08-04-012, February 3, 2010, p.44.

In theory, resource adequacy could also be tuned to achieve other reliability indexes, like system average interruption duration index (SAIDI), system average interruption frequency index (SAIFI), or their equivalent for consumers, CAIDI and CAIFI. However, such practices are not common, likely because they are less intuitive, more difficult to model on a forward-looking basis, and too focused on a single shortfall parameter. These metrics are better suited to assessment and tracking of historical system performance, and as indicators in trends of distribution system or overall system performance.

2.1.4. Translation of Resource Adequacy Standards to Reserve Margins

In its most fundamental form, the study process for a physical resource adequacy standard considers projected load profiles, installed generation capacity, production from variable generation, scheduled generator outages, and the probability of forced outages to determine the number of days in each year when a generation shortfall might occur.[18] The associated models also require the input of assumptions about load growth, load variability, energy-price responsiveness, availability of interruptible load (and DR), import capacity and availability, and potential emergency actions (e.g., voltage reduction, recalling exports, and appealing to public for conservation), among others.[19] Considering these factors, planners will simulate future scenarios, within which they will assess whether there are periods during which demand could exceed supply and, if so, how frequently. These high-risk periods will tend to be the hottest days of summer and coldest of winter, the times of year when numerous generating units are undergoing scheduled maintenance, and the seasons with low hydropower availability. However, certain circumstances in off-peak periods — for instance, if a late summer heat wave corresponded with generation outages and low hydro availability — may also turn out to be periods of expected loss of load.[20]

For each simulated period, there may be a non-zero probability that circumstances in some hours or days will result in insufficient supply and the need to shed firm load. If analysis suggests that the power system under consideration will face such conditions more often than the desired resource adequacy criterion dictates, then planners will determine how much additional capacity is needed to ensure that the criterion is met. Historically, these calculations have resulted in target capacity reserve margins between 10-20% of forecast annual peak load. These requirements then feed other mechanisms that ensure that sufficient

---

18 Systems with significant hydro capacity may place a much greater emphasis on the uncertainty associated with rainfall/drought and the amount of water that is available. Unlike systems with primarily thermal resources — said to be "peak constrained" — such systems are considered "energy constrained." See PSR, "Review of Supply Adequacy Criteria in the Northwest," prepared for Northwest Power and Conservation Council (NWPCC) and Bonneville Power Authority (BPA), September 2010, p. 3.

19 Many times, conservative assumptions are made for these values, and the result is "a very conservative criterion, conservatively applied." See supra note 11, at 19.

20 See NERC Integration of Variable Generation Task Force, "Methods to Model and Calculate Capacity Contributions of Variable Generation for Resource Adequacy Planning," March 2011, p. 10.

quantity of supply resources are actually developed in such a way that load is served at an acceptable level of reliability.

2.1.5. Reliability vs. Efficiency in Resource Adequacy

As described above, common resource adequacy standards are based on physical characteristics and outcomes rather than on economic analysis. In both their conception and execution, they often do little or nothing to consider the economic ramifications or the incentive effects associated with desired standards. On this topic, there is the most literature regarding the economic implications of 1-in-10 and, more generally, the LOLE standard.

In most critiques of resource adequacy standards, the VOLL implied by the standard is compared with a VOLL that might be considered reasonable. In theory, an efficient market would procure a level of resource adequacy to the level where the next unit of supply exceeded customers' willingness to pay for additional reliability. This level, where customers would rather go without power than fund additional capacity investment, is equivalent to the VOLL. Literature suggests that common estimates place average VOLL at between $2,000 and $5,000/MWh.[21] The situation is complicated by the fact that VOLL is difficult to measure and differs among customers and across time.[22] Some estimates can even rise into the $20,000/MWh range or higher, though customers who place the highest value on reliability (e.g., hospitals) often have arrangements that limit their curtailment in shortage situations, or have invested in alternative sources of supply (e.g., backup generators) to mitigate effects of the full range of outage situations.

The VOLL implied by the LOLE standards can be striking. For example, assuming a VOLL of $4,000/MWh, an average outage duration of five hours, and a relatively high capital cost ($120/kw-year on a net CONE basis[23]), the implied economically optimal LOLE would be about six events per year or *60* events in 10 years. Granting a higher VOLL of $20,000/MWh, the implied efficient LOLE would be 1.2 events per year, or 12 events every 10 years. Thus, even with very conservative assumptions, a 1-in-10 standard may be overly stringent by more than an order of magnitude relative to what is indicated by the economics of electricity demand.[24]

---

21 See Aaron Breidenbaugh, New York Independent System Operator, "The Market Value of Demand Response," presented at the PLMA Fall 2004 Conference, Sept. 30, 2004, or U.S. Department of Energy, "Benefits of Demand Response in Electricity Markets and Recommendations for Achieving Them," February 2006, p. 83.

22 See Andrew N. Kleit and Robert J. Michaels, "Does Competitive Electricity Require Capacity Markets? The Texas Experience," Texas Public Policy Foundation, February 2013.

23 Net CONE refers to the cost of new entry less expected revenues from the energy and ancillary services market. Generally, net CONE is the revenue that a reference unit would be expected to need to receive in addition to energy and ancillary service market revenue in order to stay in business.

24 See James F. Wilson, "Reconsidering Resource Adequacy," *Public Utilities Fortnightly*, April 2010, p.1.

In 2013, the U.S. Federal Energy Regulatory Commission (FERC) procured a study that, among other things, compared reliability-based approaches to resource adequacy to economically driven resource adequacy standards. The approach used a model purpose-built for this type of evaluation to examine a hypothetical, mid-sized Regional Transmission Organization (RTO) with reasonably realistic characteristics. Like many real-world assessments, when taking an approach based on physical reliability, the model suggested a planning reserve margin of approximately 15% to achieve 1-in-10 LOLE (events) level of reliability. However, taking an economic approach based on minimizing cost to the region and its neighbours, the study revealed that a planning reserve margin of approximately 8% — across the primary RTO and its neighbours — would be efficient. Moreover, efficient reserve margins would be even lower for a smaller system or a system with well-interconnected neighbours who were long on installed generation.[25]

The FERC study also sheds light on the magnitude of the financial trade-off between reliability- and cost-based approaches to resource adequacy. As described above, reliability-based approaches dictate procuring more supply than would be economically efficient if based on the value consumers are inferred to place on reliability. However, in over-procuring capacity, the system is able to reduce energy price volatility and minimize the frequency of expensive, extreme events, while also dramatically reducing the frequency of supply shortages and interruptions. Such an approach has clear political advantages for system operators and policymakers. Modeling in the FERC study suggests that increasing the planning reserve margin for the sample system from the cost-minimizing level of 8% to the 1-in-10 level of 15% only increases customer costs by 1.5%.[26] While this may be a significant sum in absolute terms, it is relatively small compared to overall system costs; this trade-off comes with the aforementioned benefits that policymakers have historically found attractive. However, opting for a system that dictates holding capacity in excess of consumers' willingness to pay ultimately increases supply available in the spot market and suppresses energy and ancillary service market prices below levels that would support investment to the desired reliability level. In turn, such a system places increased reliance on a capacity market or other mechanisms to ensure there is sufficient revenue available to encourage investment in new generation and in continued operation of existing resources.

2.1.6. Alberta's Approach to Resource Adequacy

Alberta does not currently mandate a physical resource adequacy standard. Under Alberta's current energy-only market, the level of adequacy over the long term is a result of decisions by private investors responding to market signals.

25 See Johannes P. Pfeifenberger, *et al*., "Resource Adequacy Requirements: Reliability and Economic Implications," The Brattle Group, Astrape Consulting, September 2013, p.vi.

26 *Id.* p. viii.

The AESO addresses resource adequacy indirectly by creating a "bridging mechanism" in the event that adequacy becomes an issue during a two-year forecast period, and action has to be taken to maintain adequacy until new capacity is built or load decreases.[27] What qualifies as a level of concern that justifies action is defined in ISO Rule 202.6. The AESO monitors and reports on long-term adequacy on a quarterly basis through four metrics: new generation status and retirements, a forecast reserve margin metric, a supply cushion metric, and an estimation of the two-year probability of supply adequacy shortfall metric. The two-year probability of supply adequacy shortfall metric is used as a threshold test to determine if preventative action is required to maintain resource adequacy until new capacity is built or load decreases. Specifically, if the metric indicates that on a probabilistic basis, unserved energy exceeds 1,600 MWh in any two-year period, the AESO may procure one or more of load-shed capability, backup generation or emergency portable generation services.[28] The threshold actions are intended to have minimal market effects in that they would be called upon only when the price rises to the price cap.

2.2. Economic Rationales for Capacity Markets

This section explores the various economic rationales for the need for capacity markets in competitive wholesale electricity markets. Capacity markets have emerged as a means to solve the resource adequacy problem, or what has also been referred to as the "missing money" problem.[29] The missing money problem arises when the expected net revenues from sales of energy and ancillary services provide inadequate incentives for merchant investors in generating capacity, or equivalent demand-side resources, to invest in capacity sufficient to meet established resource adequacy standards.[30]

In addition to adherence to physical resource adequacy standards, there are various market and regulatory failures that distort investment signals and contribute to missing money. The following provides a brief summary of its theoretical and practical causes.

2.2.1. Gaps Between Economic Reserve Margins and Required Reserve Margins

In theory, a well-designed energy-only market will, over time, support a certain level of resource adequacy and an associated economic reserve margin above peak demand levels. However, as discussed above, economic analysis based on reasonable estimates of average VOLL and the CONE implies that this

27 See https://www.aeso.ca/assets/downloads/Long-Term-Adequacy-Feb-7.pdf at p.4.

28 See https://www.aeso.ca/rules-standards-and-tariff/iso-rules/section-202-6-adequacy-of-supply/.

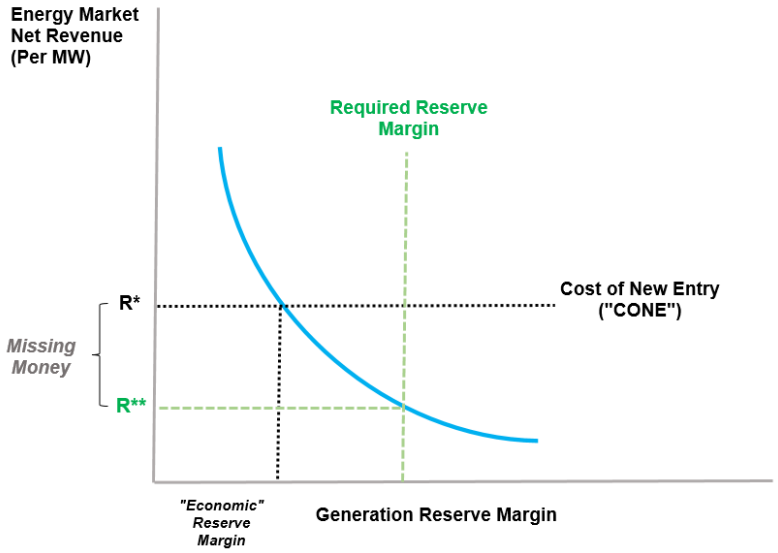
29 The issue was first termed the "missing money" problem by Roy Shanker, Comments on Standard Market Design: Resource Adequacy Requirement. Federal Energy Regulatory Commission, Docket RM01-12-000.

30 See supra note 9. Joskow also refers to the "missing money" problem as the "revenue adequacy" problem. Note that not all jurisdictions that operate energy-only markets have suffered a missing money problem. Sections 3.2 and 3.3 discuss how jurisdictions such as ERCOT and Alberta have addressed the problem through means other than capacity mechanisms.

economic reserve margin is generally lower than the reserve margins required to support the traditional resource adequacy standard. Mandating these requirements can induce the missing money problem. This is illustrated in Figure 1 below.[31]

Figure 1 plots the expected relationship between energy market net revenues ($/MW) earned by a benchmark generator and the system generation reserve margin (percent of installed capacity). The lower the generation reserve margin (i.e., the lower the amount of installed generation capacity), the higher the energy market prices and the associated net energy revenues earned for a benchmark generator. That is, at lower levels of installed generation capacity, there are more instances in which there is a relative scarcity of energy supply, causing more instances of higher scarcity energy prices and associated energy margins for a benchmark generator.

Figure 1: Economic Reserve Margins and Missing Money



In theory, a well-designed and perfectly functioning energy-only market would attract the optimal, social welfare-maximizing level of investment. In such a market, prices would reflect the marginal cost of generation and the marginal benefit of consuming as defined through the intersection of supply and demand. Implicit in this well-designed market is the assumption that consumers can choose how much they are willing to pay for reliable service.[32] As available generation supply becomes relatively scarce, prices rise and customers that place a lower value on reliability choose not to consume. With more extreme levels

31 Figure 1 is adapted from Robert Stoddard and Seabron Adamson, "Comparing Capacity Market and Payment Designs for Ensuring Supply Adequacy," Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009.

32 This description of a perfectly functioning market also assumes consumers are aware of the price increase as it happens in real time. The current impracticality of this assumption is the focus of the discussion in Section 2.2.2.

of generation scarcity and capacity constraints, market prices rise and reflect the marginal benefit of consumers that place a higher value on reliability.

In this well-functioning market, prices always reflect the marginal social cost and marginal social benefit of the actions of suppliers and consumers. In the long run, these prices induce new entry to the point where the net energy revenues earned by the benchmark generator are just enough to recover fixed operating costs and the cost of entry. This produces an economic reserve margin as highlighted in Figure 1 with a level of net energy market revenues equal to R*.[33]

The missing money problem can arise when policymakers place a higher value on reliability than consumers, and choose a reliability standard that requires the system operator to maintain a generation reserve margin that exceeds the economic reserve margin R*. This is illustrated in Figure 1. A higher required reserve margin, and hence, more installed generation capacity on the system, means lower net energy market revenues for all generators, represented as R** in Figure 1. The difference between R* and R** represents the missing money, or the additional revenue, which may be expressed on a per megawatt basis, required by generators to support the private investment needed to achieve the desired level of reliability. The gap between the economic reserve margin and the required reserve margin provides the justification for a capacity mechanism to induce private investment to the required levels.[34]

2.2.2. Demand-Side Market Imperfections

To some, the explanation of the missing money problem as described above would appear to be caused by regulatory or policy intervention, the solution to which would simply be to minimize the role of the regulator and to allow the market to solve for reliability. However, Cramton, Ockenfels and Stoft have argued that as a result of certain demand-side market flaws, a wholesale energy market cannot operate satisfactorily on its own to achieve the economically optimal reserve margin.[35] Instead, the market requires administrative intervention.

The hypothetical, well-designed electricity market described above always clears. That is, there always exists a market price wherein demand and supply are in balance. However, as Cramton et al. point out, in

33 In a well-designed, energy-only market, the net energy market revenues for all generation types (baseload, intermediate and peaking generation) should equal their investment costs or exit costs in equilibrium.

34 This discussion focuses on a static analysis of the market and market failures that motivate a capacity mechanism to induce investment. When one adds additional considerations to the analysis, like uncertainty, market dynamics, and the "lumpy" nature of investments, the conclusions do not change. If anything, a consideration of broader issues is likely to exacerbate market failures.

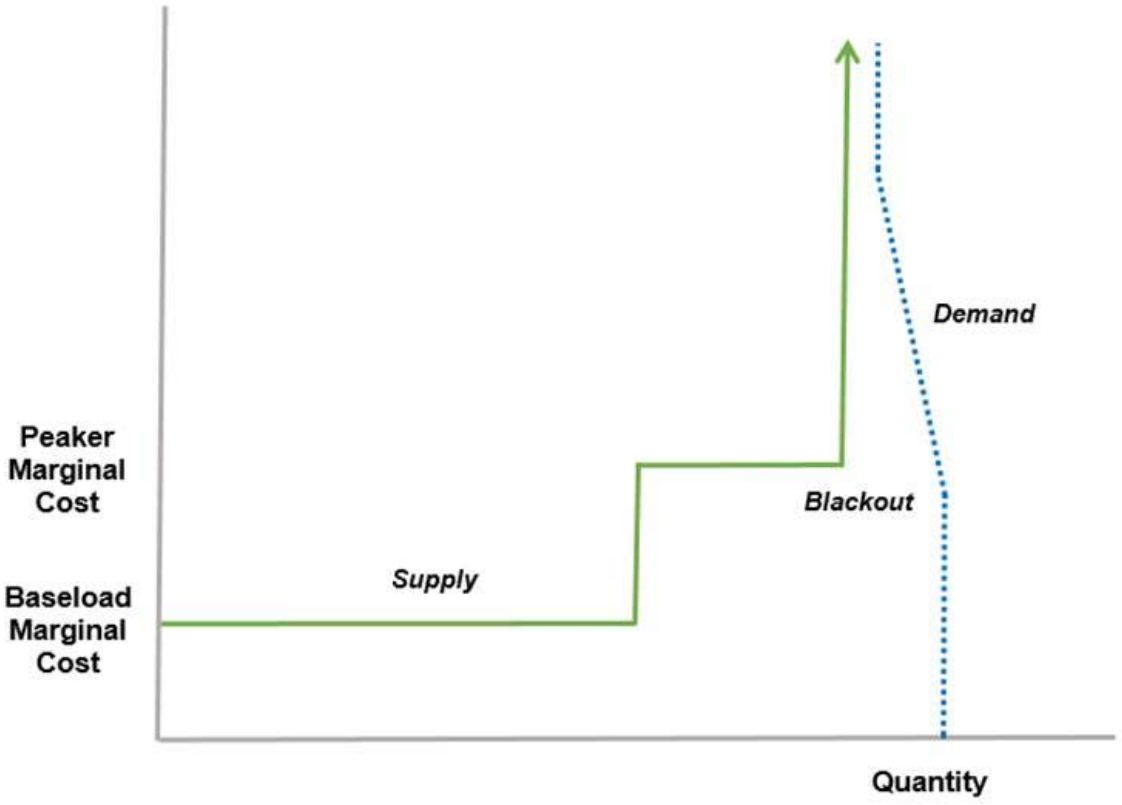
35 This section closely follows Peter Cramton, et al., "Capacity Market Fundamentals," May 26, 2013. *See also* Peter Cramton and Steven Stoft, 2006, "The Convergence of Market Designs for Adequate Generating Capacity," white paper for the California Electricity Oversight Board, March 2006, and Peter Cramton and Axel Ockenfels, "Economics and Design of Capacity Markets for the Power Sector," 2011.

practice, current electricity markets are not always guaranteed to clear. The main reason for this is low demand flexibility. In particular, the lack of real-time meters and billing and other equipment to allow consumers to see and respond to real-time prices, means that demand eventually becomes perfectly inelastic in real time.[36] At the same time, given that it is currently costly to store electricity, the supply-side also eventually becomes perfectly inelastic. This creates the potential for demand to exceed supply. In this situation, there is no price at which the market can clear. With insufficient demand flexibility, there is a possibility of involuntary load reduction, or brownouts, if generation capacity is not adequate. This is illustrated in Figure 2.

The dotted line presented in Figure 2 depicts the demand curve for energy. It is perfectly inelastic at lower price levels, exhibits some price elasticity at medium price levels, and eventually becomes perfectly inelastic at higher price levels. The stepped curve in Figure 2 depicts a stylized version of the energy supply curve. It, too, becomes perfectly inelastic as the available generation capability is exhausted. In Figure 2, the energy demanded is greater at all price levels than what the supply curve indicates is available. The demand for energy exceeds the capacity of generation available to produce energy. In this situation, there is no price at which demand and supply intersect. As a result, there is no equilibrating price in the market.

36 The lack of real-time meters and the inability of consumers to respond in real time would not be fatal if LSEs or the system operator could contract in advance with consumers to prioritize customer rationing according to their individual preferences for reliability. See H.P. Chow and R. Wilson, "Priority Service: Pricing, Investment and Market Organization," *American Economic Review* 77, 1987, pp. 89–116. However, the system operator cannot control the flows of power to individual customers so as to enforce such contracts.

Figure 2: Absence of Sufficient Demand Response



When the market does not clear, the price paid to generators must be set by administrative pricing rules. The system operator must use non-price rationing of demand to maintain the balance of supply and demand.

In this situation, given imperfect information on the VOLL of individual consumers, the administrative price rules cannot solve for the optimal level of reliability and consequently, the optimal economic reserve margin. Administrative prices that are too high result in too much capacity (too few interruptions) and administrative prices that are too low — the more common case — result in too little capacity (too many interruptions). At best, if the system operator can assume a reasonably accurate average VOLL, and can set the administrative price at this level, the market may solve for the "optimal" second best generation reserve margin.[37] However, this assumes that the market is well designed in all other ways, which, as discussed more below, is not always the case. Cramton argues that this provides a rationale for capacity markets — "to provide the amount of capacity that optimizes the duration of [brownouts]."

37 This is a second best optimum as individual consumers have different VOLL, and there is no assurance that the non-price rationing has selected the appropriate consumers (i.e., those that have a relatively lower VOLL).

2.2.3. Operating Reserves as a "Public Good"

Another argument for administrative intervention due to market imperfections is derived from the view that operating reserves have attributes of a public good.[38] Operating reserve is generation capacity (or in some markets, load reduction) that is held on standby, and can be called on by the system operator to produce (reduce) energy in a short interval of time (10 minutes to 30 minutes) in response to a sudden grid contingency such as the outage of a generating plant or transmission line. Operating reserves are a regular feature of all electricity markets, including those referred to as energy-only markets. Failure of a market design to price operating reserve properly can contribute to the missing money problem.

Operating reserves serve two purposes. First, operating reserves reduce the probability of non-price rationing events (i.e., controlled load shedding) due to a sudden contingency on the system. Second, operating reserves are used to ensure system security — operating reserves respond sufficiently quickly to a grid contingency so as to maintain the frequency, voltage, and general stability parameters of the network — and prevent cascading failures and network collapse.

During a network collapse, there is no market price. Generators that are available to produce, but are rendered idle by the collapse, are unable to earn revenues. Generators cannot profit from the overall scarcity of energy during a collapse. In this regard, all available generators benefit from the provision of operating reserves as it prevents a network collapse and a sudden unexpected loss of revenues for all generators. In the hypothetical, to the extent that a single generator would choose to provide operating reserve (place its own available assets on standby rather than produce energy for sale) to protect against a system collapse, all other generators would have the incentive to free-ride on this single generator's decision — the single generator could not exclude generators that refused to pay from benefiting from its provision of reserve.

Joskow and Tirole argue that this makes operating reserves a public good, and without centrally mandated operating reserve requirements and procurement of reserve by the system operator, there would be under-provision of operating reserves and lower reliability than is optimal. Furthermore, Joskow and Tirole note that "under certain contingencies, the market price and the associated scarcity rents available to support investments in generating capacity are extremely sensitive to small mistakes or discretionary actions by the system operator."[39] These small mistakes and discretionary actions are likely to lead to the market prices undervaluing the operating reserves in these instances, particularly during periods of scarcity, and consequently, contribute to the overall missing money problem.

---

38 See Paul Joskow and Jean Tirole, "Reliability and Competitive Electricity Markets." *RAND Journal of Economics*, 38(1), 2007, p. 82.

39 *Id.* at p. 83.

2.2.4. System Operator Reliability Procedures During Scarcity Events

There is also evidence from some U.S. markets that system operators' administrative procedures used to manage shortage conditions and potential non-price rationing of demand can inefficiently depress market prices during periods of scarcity, further contributing to the missing money problem.[40] These represent energy market failures driven by the system operators' mandate to manage reliability — the effects of which will vary across system operators based on their individual interpretation of their mandate.

For example, system operators' decisions to implement voltage reductions to reduce demand, just prior to involuntary load reductions to stabilize the system, has the effect of reducing prices relative to the price level at normal voltage. This implies a downward bias of market price signals during these events. Similarly, in anticipation of a pending generation shortage caused by a contingency or sudden and unexpected increase in forecast demand, system operators may call on generators out of market to guard against the use of voltage reductions. Such actions are particularly problematic from an efficiency perspective when the system operator calls on reserves and provides compensation at the *ex ante* procurement cost of reserve, which is likely to be much lower than the scarcity price.[41] This, too, has the effect of artificially lowering prices during relative scarcity events.[42] In both cases, these events typically persist for a short duration and represent second order drivers of the missing money problem.

2.2.5. Regulatory Responses to Market Power

The conditions that can lead to scarcity prices (inelastic demand and supply) are also the conditions that are conducive to the unilateral exercise of market power. In fact, it is often difficult to distinguish between high prices caused by scarcity and high prices that result from the exercise of market power. The concern over the exercise of market power and the political attention paid to high prices has drawn a response from regulators to impose price caps that are well below the VOLL, and to impose generator offer caps or automatic offer mitigation procedures that reflect engineering notions of marginal cost that can ignore full

40 See Paul Joskow, "Capacity Payments in Imperfect Electricity Markets: Need and Design," *Utilities Policy* 16, 2008.

41 This particular issue may be addressed by implementing an Operating Reserves Demand Curve construct, which is described in more detail in Section 3.2.

42 See William H. Hogan, "Connecting Reliability Standards and Electricity Markets," Harvard Electricity Policy Group, Presentation, December 8, 2005.

incremental costs or actual opportunity cost.[43] These measures can collectively depress market prices below those that truly reflect competitive supply, leading to missing money and underinvestment.[44]

2.2.6. Price Volatility and Illiquid Forward Markets

Cramton and Ockenfels argue that there are additional reasons for why capacity markets are useful.[45] Due to inelastic demand and relatively inelastic supply, prices in electricity markets can be extremely volatile and hard to predict. Hours during which the capacity is near full utilization, and net revenues rise to levels sufficient to cover (or exceed) investment costs, can be rare. Years can pass during which average prices do not allow investors to recover annual costs, while just a few hot days during a single summer can create high enough spike prices for them to profit significantly and make up for years of losses. The magnitude and frequency of price spikes, or lack thereof, can be magnified further by boom and bust development cycles, the "lumpy" nature of generation investment and regulatory interventions. The fluctuating nature and high risk of electricity prices can be discouraging to developers that might invest in new capacity or upgrade existing facilities. However, investment in new capacity is necessary to meet increasing demand, or to replace older units becoming less efficient and eventually retiring, or even to defer the retirement of old units.

Well-functioning forward markets could provide some ability to hedge fluctuations through long-term contracts. However, forward markets for electric energy are often illiquid, with few participants available for the required contracts to secure adequate supply for customers. Long-term contract prices may also vary considerably depending on circumstances both in and out of the market, further increasing the difficulty for market participants to arrange for consistent returns on costs. Capacity pricing mechanisms hold more promise. Though even a well-functioning capacity market cannot eliminate price volatility, it can mitigate risk to market participants by replacing peak energy rents with a more predictable overall payment stream predetermined for a commitment period. Despite continued volatility in the spot market during times of scarcity, investors may be more willing to build in a regime where there is additional opportunity to earn revenue, particularly if accompanied by some degree of increased certainty, in which case, overall revenue from both the capacity and energy markets is more likely to fully cover costs.[46]

---

43 Approaches to energy offer mitigation in U.S. markets is discussed in more detail in Section 4.7.3.

44 See William H. Hogan, "Connecting Reliability Standards and Electricity Markets," Harvard Electricity Policy Group, Presentation, December 8, 2005, and Peter Cramton and Axel Ockenfels, "Economics and Design of Capacity Markets for the Power Sector," 2011, available at ftp://www.cramton.umd.edu/papers2010-2014/cramton-ockenfels-economics-and-design-of-capacity-markets.pdf.

45 See Peter Cramton and Axel Ockenfels, "Economics and Design of Capacity Markets for the Power Sector," 2011, pp. 12–13.

46 Ibid.

Note that while the lack of a well-functioning, liquid-forward market may contribute to the resource adequacy problem, a well-functioning, liquid-forward market on its own will not address the various market and regulatory imperfections (discussed above) that a capacity market is intended to address. Though forward contracts mitigate uncertainty from price volatility, they do not affect those spot prices that cause the missing money in the first place. Through the pressures of intertemporal arbitrage, buyers and sellers will only agree to forward contract prices that reflect the expected future spot prices. To the extent that these spot prices suffer from the missing money problem, the forward contract prices will be insufficient to incentivize capacity to the levels required to meet the established resource adequacy standards.[47] Furthermore, Cramton and Ockenfels argue that long-term, forward, energy-only contracts can aggravate resource adequacy if such contracts are used to attract excessive resources that cause spot prices to fall, thereby reducing the value to the prior investments of incumbent generators not party to the contracts.[48]

2.2.7. Variable Generation and Other Subsidized Capacity

Subsidized generation, particularly subsidized variable generation, has the potential to exacerbate the missing money problem for other generators. Many jurisdictions with developed markets for electric power also have renewable energy policies, in the form of subsidies or mandates. These jurisdictions are experiencing rapid growth in installed renewable capacity, much of it in wind and solar facilities with variable production capability.[49] Though it is not always windy or sunny, the general growth in available capacity drives down energy market prices below the efficient levels that would arise absent such policy intervention. At the same time, the availability of low (or zero) variable cost generation reduces the utilization of legacy resources. The overall effect can be to reduce revenues dramatically to other generators. The economically efficient outcome, then, is for much of the legacy fleet to retire to the extent that it is unable to recover its costs through the marketplace under the new supply mix. In this circumstance, the missing money problem may be more relevant for existing plants, and system operators may face concerns over ensuring that price signals are appropriate to avoid having too much generation exit the market (as opposed to the frequent concern over underinvestment in new generation) such that reliability becomes an issue.

Furthermore, the specific incentives faced by renewables to offer and operate in the market can distort static dispatch efficiency and, possibly, the long-run efficient responses of investors in an energy-only market. For example, contracting approaches that are “take-or-pay” — i.e., pay a fixed price for energy produced regardless of what the actual market price is — can incentivize inefficient production during

47 Id. p. 16. *See also* Paul Joskow, “Competitive Electricity Markets and Investment in New Generating Capacity,” *The New Energy Paradigm* (Dieter Helm, ed.), Oxford University Press, 2007, p. 49.

48 *See supra* note 45, at 16.

49 Under the CLP Alberta plans to have as much as 5,000 MW of renewable capacity, which would represent a substantial increase from current levels and would most likely materially affect energy prices.

periods of low demand when there is a surplus of other low marginal cost resources operating. During these surplus conditions, the "take-or-pay" contract can incent the renewable producers to continue to produce when it is socially inefficient to do so. For example, it may be more costly to shut down a thermal unit for brief intervals in order to manage a surplus of supply than to dispatch down a wind generator. Moreover, by shutting down the thermal unit, it may become unavailable in future periods, in which case more expensive generation may be required.[50]

2.2.8. Transitional Mechanisms in Support of Major Policy Change

Wholesale electricity markets may face periods of interruption in which exogenous factors temporarily undermine the efficient operation of energy and ancillary service markets. Of particular relevance at present, many power systems are in the midst of policy-driven transitions from heavy reliance on traditional thermal generation to supply mixes that are substantially more focused on clean energy sources. During these transitions, there are often large quantities of new, subsidized generation built alongside existing supply sources. As described in Section 2.2.7, there may be temporary circumstances where installed capacity far exceeds what would exist in an efficient equilibrium, with capacity supply gluts leading to suppressed prices and exacerbation of the missing money problem. While economics will eventually drive legacy thermal units to retirement, there may be a part of the transition period in which some such units remain important for reliability purposes. To ensure a smooth transition in which reliability is not threatened by mass retirements, a temporary mechanism to provide additional revenue to resources — not necessarily a capacity market — may be implemented until such a point as the system has had time to settle into a new policy paradigm. Temporary measures may include any mechanism described in Section 3, or other approaches such as reliability must-run contracts. Simpler measures may be advisable in the transitional case so as to limit the effort expended to develop a temporary mechanism.[51]

Major periods of transition are often also accompanied by regulatory uncertainty. This can lead to an increase in perceived risk, higher costs of capital, and reticence on the part of developers to construct new supply resources. Temporary mechanisms that provide certainty to investors, possibly by compensating suppliers for the provision of capacity, may be an attractive tool to help mitigate the effects of regulatory uncertainty. However, there is skepticism in the literature as to whether capacity mechanisms, particularly capacity markets, are an efficient solution to problems caused by the effects of policy on markets. Cramton et al. suggest that the capacity market is the wrong tool for facilitating transitions and that, "It is desirable to

50 Alberta also has the potential to expand its use of hydroelectric generation in the province through the Renewable Electricity Program. See http://www.energy.alberta.ca/Electricity/pdfs/HydroelectricPowerInquiry.pdf at p. 3. The importance of properly designed contracts to incentivize efficient offers into the energy market are particularly apt for energy-limited hydroelectric resources.

51 "Capacity Mechanisms in Individual Markets Within the IEM," Thema Consulting Group, June 2013, p. 46, available at https://ec.europa.eu/energy/sites/ener/files/documents/20130207_generation_adequacy_study.pdf.

firmly address [transitional] issues before a capacity market is adopted. No capacity market can function well if there are impediments to long-term investment, such as political uncertainties, regulatory imperfections causing poor implementation, insufficient development of locational, and real-time pricing, etc."[52]

2.2.9. Summary of Rationales for Capacity Markets

In short, capacity markets have emerged as a means to address the resource adequacy problem or what is more often referred to as the "missing money" problem. The missing money problem arises when the expected net revenues from sales of energy and ancillary services earned at market prices provide inadequate incentives for merchant investors to generate capacity, or equivalent demand-side resources, to invest in sufficient capacity to meet established resource adequacy standards.

There are a number of possible factors that contribute to the missing money problem in energy-only markets, including the pursuit of physical rather than economic reliability standards, wholesale market imperfections or market failures, regulatory constraints on market prices and generation bid parameters, and system operator procedures for managing their operational reliability mandate. Collectively, these factors act to suppress real-time, wholesale energy and operating reserve prices below efficient levels and below those needed to attract sufficient capacity to meet resource adequacy without further intervention. Adding a capacity market to an energy market is one approach to addressing the missing money problem by providing an additional revenue stream to supply resources, although there are other approaches presented in the academic literature or implemented in practice. The next section provides a brief description of the various approaches.

3. Approaches to Addressing the "Missing Money" Problem

There is no universally agreed-upon approach amongst scholars and practitioners on how to address the resource adequacy or missing money problems. A variety of approaches have been presented in the literature or implemented in practice, each with its own pros and cons. This section provides a brief summary of these approaches. The approaches can generally be described as falling within two categories. The first is capacity mechanisms, which essentially split payments to generators into two, one payment for the spot market sales (energy and ancillary services) provided, and another payment for a separate product,

52 Peter Cramton and Axel Ockenfels, "Economics and Design of Capacity Markets for the Power Sector," 2011, p. 3. Alberta does not have locational, real-time pricing but instead operates a uniform energy market and has a "congestion-free" transmission policy. The implications of this for the introduction of a capacity market is discussed more in Section 4.3.

capacity. The second is the energy-only approach, which focuses on raising the price of energy during periods of scarcity.

3.1. Capacity Mechanisms

Capacity-based approaches to addressing resource adequacy divide the electricity market into two products, energy and capacity, the latter representing the availability of capacity to produce energy, particularly during periods of scarcity or peak demand. There are three main capacity mechanisms: capacity payments, strategic reserves and capacity markets, the last of which includes decentralized bilateral capacity markets, centralized installed capacity markets and reliability options.

3.1.1. Capacity Payments

Perhaps the simplest approach is to provide generators with a direct "capacity" payment to supplement the revenue earned through the energy and ancillary services markets. This approach was used most extensively in the early designs of many European energy markets.[53]

Capacity payments are generally established by a regulatory body as a payment to strengthen incentives for investing in new capacity or old capacity.[54] There are several ways to calculate the capacity payment. However, the general intention is to provide a payment equal to the missing money.

Capacity payments may be market-wide, applying to all capacity operating in the market, or designed as targeted payments to new or peaking generation. With a capacity payment, there is typically no explicit reliability standard or reserve margin obligation imposed on load serving entities (LSE) or the system operator. However, the level of the payment may be calibrated against system and economic metrics like CONE, VOLL and/or the LOLP. The capacity payments are recovered from consumers by their LSEs with the approaches to cost recovery varying across entities.

There are several weaknesses to this approach, a factor leading to their abandonment in favor of other mechanisms in most markets. First, it is difficult to determine the right amount of payment to be made *ex ante,* and the payments may have no or little relationship to actual supply and demand factors. As such, they do not provide good economic signals for investment. Second, there is generally no clearly defined capacity product, so generators are paid regardless of their contribution to overall system adequacy and reliability. Third, target capacity payments that only compensate select technologies or new generation can often lead to the under compensation of existing generation, causing inefficient exit. Finally, as the

53 For a brief review, see Robert Stoddard and Seabron Adamson, "Comparing Capacity Market and Payment Designs for Ensuring Supply Adequacy," Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009. This approach was used in the early designs of the UK market and Irish market. Other examples include Spain, Greece, Chile, Colombia and Peru.

54 See https://ec.europa.eu/energy/sites/ener/files/documents/20130207_generation_adequacy_study.pdf.

payments are not tied to a reserve margin target, there is no guarantee the payments will drive the level of reliability needed to meet the standards, the very objective to which the payments are designed.

3.1.2. Strategic Reserves

A second, fairly simple capacity mechanism is the procurement of strategic reserves. Strategic reserves are used in several European markets such as Belgium, Denmark, Germany, Poland and Sweden.[55] Strategic reserve is capacity that is on standby to provide energy when called upon by the system operator. The strategic reserve capacity does not participate in the energy market.

Strategic reserves are implemented through the imposition of a reliability obligation on the system operator by a regulatory authority to hold a specific amount of capacity based on a forward-looking study of resource adequacy.[56] The system operator then holds a tendering process for the fixed amount of capacity. The tender may be directed at specific technologies (e.g., peakers), may be available to existing capacity, and may be open to DR resources. The winners of the tender then sign contracts with the system operator that specify payment structures and provisions around notification, activation and compensation during activation. The costs of the strategic reserves are generally recovered from consumers as a system charge included in the transmission tariffs.

Strategic reserves are typically activated in one of two ways. The activation may be triggered when energy prices hit the price cap. Alternatively, the activation may be triggered when there is a supply shortfall in the day-ahead market. In this latter situation, the energy price is administratively determined.[57]

The appeal of strategic reserves is that they are fairly simple to implement and operate. However, they are unlikely to promote the most efficient means of achieving resource adequacy. First, when activated, the energy price is capped, which can aggravate the missing money problem for generators participating in the energy market if the cap is lower than average VOLL. Second, strategic reserves can undermine investor confidence if there is a belief that they will be activated for political reasons (e.g., to lower prices and consumer costs). This could also lead to the risk that existing plants would threaten to close unless compensated as strategic reserve. Third, strategic reserves are likely to result in less efficient dispatch than other capacity mechanisms. The resources acting as strategic reserves are held out of the energy market until triggered, even when their operating cost may be cheaper than the cost of other generators operating in the energy market. Finally, the tenders for strategic reserves are typically targeted to a

55 See https://ec.europa.eu/energy/sites/ener/files/documents/swd_2016_385_f1_other_staff_working_paper_en_v3_p1_870001.pdf.

56 Strategic reserves are similar in nature to operating reserve except that they are procured for resource adequacy reasons rather than for system adequacy, and well in advance of real time and for a longer term.

57 See https://ec.europa.eu/energy/sites/ener/files/documents/20130207_generation_adequacy_study.pdf.

particular technology or to new generation, limiting the possibility that other generation types might compete by offering to provide services at a lower price.

3.1.3. Decentralized Bilateral Capacity Markets

Under this capacity mechanism, a capacity obligation is placed on a market participant, such as an LSE, to contract bilaterally with capacity providers to secure the amount of capacity needed to meet their share of the overall resource adequacy needs of the system. California's resource adequacy program is based on decentralized capacity obligations.[58]

In a decentralized capacity obligation mechanism, a central regulatory or government authority establishes the resource adequacy standard, and in conjunction with the system operator, determines the associated reserve margins for the overall system. Each LSE is then assigned an annual obligation to procure enough capacity to meet the peak demands of their customers plus their share of the system reserve requirement (generally stated as a percentage of their peak demand).

LSEs can then fulfill their obligations through the ownership of plants (or qualified DR programs) or through bilateral contacts with power generators. LSEs may also trade capacity resources with other LSEs in order to meet their obligations, leading to the emergence of a bilateral market for capacity. LSEs pay the generators according to the terms of the contract. LSEs and generators may see an interest in signing a longer-term contract for all or parts of the LSE's future needs or choose shorter-term "spot" arrangements. LSEs recover the cost of the capacity contracts from their end-use consumers through retail rates, either as an energy charge or a demand-based charge.

An LSE must demonstrate to the system operator that it has fulfilled its obligations and reference the specific resources from which it has procured the capacity to assure there is no double counting. LSEs that fail to fulfill their obligations are subject to a fine, typically equal to the cost to the system operator to replace the capacity plus a penalty.

Generally, generation resources that are counted toward an LSE's obligation must offer into the day-ahead and real-time markets, with some allowance made for outages. Performance penalties may be applied to generators to incentivize their availability.

An advantage of a decentralized obligation model is that it does not require a complex auction design process with centrally determined designed parameters (discussed more below). Through bilateral trading, the cost of capacity should reflect the supply and demand fundamentals with capacity prices falling when there is an oversupply of capacity and rising when there is an undersupply. As an added benefit, the

58 Decentralized obligations also exist within other U.S. jurisdictions such as PJM, NYISO and MISO, although these jurisdictions also operate centralized capacity markets as discussed further below.

resulting bilateral arrangements hold the potential to allow for efficient sharing of short-term versus long-term risk between LSEs and generators. Furthermore, a decentralized obligation model can be structured in a way that the nature of the capacity contracts is not administratively prescribed, which can, in turn, provide flexibility in the procurement structure and open the door for innovative approaches to fulfilling capacity obligations.

On the other hand, the approach does require administrative intervention and the development of complex rules and procedures. This drawback is also a feature of centralized capacity markets, though a centralized capacity market is arguably *more* administratively complex than what is required to set up a scheme for decentralized capacity obligations. Decentralized bilateral capacity markets are also less transparent than centralized capacity markets. Additionally, with decentralized capacity obligations, if the penalty for LSEs for failing to meet their obligation is too low, it may lead LSEs to opt to default on their obligation and subsequently risk underinvestment in capacity.

3.1.4. Centralized Installed Capacity Markets

Similar to the decentralized bilateral capacity market, the centralized installed capacity market relies on competitive forces to ensure supply is sufficient to meet forecast future peak-demand levels plus a reserve margin. However, the centralized installed-capacity market is implemented through an auction process run by the system operator (with appropriate regulatory notice and approvals), which determines the competitive price for capacity based on expected future demand and offers from qualified supply. The auction-clearing price and the resulting revenues are expected to reflect the level of missing money required by the marginal capacity resource to ensure that it covers its costs. In conjunction with the energy and ancillary service markets, capacity markets complete the suite of available economic signals needed to attract private investment. The centralized, installed capacity market approach is used by the New England ISO (ISO-NE), in the PJM Interconnection (PJM), the New York ISO (NYISO) and Midcontinent ISO (MISO).

Qualifying providers, which can include DR participants, compete against each other to sell capacity in the market. The capacity product should be clearly defined, ensuring that all MW procured are providing the same reliability benefit to the system. In addition to qualification rules that require demonstration of capability to deliver the capacity product, performance requirements (paired with incentives and punishments) can ensure that resources actually provide reliability during periods of system stress. Because all suppliers with a capacity obligation are expected to be providing the same service, capacity prices resulting from the auction are generally paid to all market participants.

Supply in the capacity market auction is relatively straightforward and easy to conceptualize. A supply curve can be assembled from offers by qualifying participants. A demand curve is constructed based on administrative rules that consider both how much capacity is expected to be needed and in what price range. The quantity components of the demand curve are based on a determination, usually by the market operator, of the installed reserve margin necessary to fulfill the relevant resource adequacy criteria. The

price components are based on the expected CONE as well as expected energy and ancillary service revenues, used together to calculate net CONE, which provides a conservative representation of the expected capacity market revenue sufficient to compensate for missing money from the other markets.

Capacity markets have the advantage of employing market forces to establish an (arguably[59]) efficient price necessary to maintain sufficient supply to satisfy an administrative reliability standard. A centralized market is also relatively transparent, facilitates a level playing field across resource types, provides a single venue in which market power can be monitored, and ensures backstop procurement of a minimum acceptable level of resource adequacy. Moreover, since the level of procurement is directly tied to the target reserve margin, with the use of a sloped demand curve, the auction may also reflect the declining (but non-zero) value of procuring additional capacity beyond the minimum required to meet the resource adequacy standard. [60]

Centralized capacity markets also have disadvantages. Above all, such markets are based on exogenously imposed requirements; LSEs/consumers would not participate in capacity markets absent the mandated requirement to procure capacity. This, in turn, means that the market design must be administrative, and will necessarily suffer from all of the associated weaknesses. The market design will derive from the cumulative contribution of, and interactions between, stakeholders, all of which are pursuing their individual interests. The literature suggests that significant imperfections are likely to derive from planned markets.[61] Not only is administrative process a challenge of the basic market design, but protests are likely to persist with ongoing administrative determinations, such as demand forecasting and estimations of CONE, that affect market outcomes. In addition to the ramifications of administrative market design, there are challenges associated with representing demand (as with all electricity markets), the need for additional layers of market rules to ensure market performance in real time, management of market power, and the treatment of alternative resource types like DR, renewables and interconnectors.

Further discussion of basic elements of centralized capacity markets is provided in Section 4.

---

59 Whether a capacity market price is truly efficient will always be a matter of debate. The fundamentally administrative nature of the capacity market leaves it prone to questions about whether the overall market design — including energy and ancillary services markets — is actually determining an efficient price and compensation scheme. There will always be factors that affect market outcomes that are fundamentally negotiated, like the appropriate values for CONE or net CONE.

60 See Johannes Pfeifenberger and Kathleen Spees, “Best Practices in Resource Adequacy,” presentation January 27, 2010, slides 10-12.

61 See supra note 22, at. 9.

3.1.5. Reliability Options

The reliability options (RO) approach to ensuring resource adequacy is a variation on centralized capacity markets that relies on financial incentives rather than physical obligations to support the availability of energy or ancillary services when needed. In this mechanism, the system operator runs a centralized auction for ROs, which are call option contracts where the holder of the option is paid an annual payment in return for the system operator having the right to call on the option holder to provide energy at a predetermined strike price. This means that the holders of ROs are effectively required to rebate to the system operator the difference between the relevant market price and the strike price when the market price exceeds the strike price. The benefits of such a scheme are realized on both the supply and demand side. On the supply side, entities that sell reliability options opt to forgo prices above the strike price during peak periods in exchange for the certain revenues of selling an RO. On the demand side, consumers pay to purchase the option and in return, receive a hedge against prices above the strike price. Due to the close linkage of the capacity mechanism to the energy price, an RO construct requires a well-functioning energy spot market and market price that acts as a reliable reference for the establishment and activation of the strike price.[62]

ROs are designed to provide signals to investors to build appropriate types of capacity to fulfill their financial obligations to the system operator during periods when the market price exceeds the strike price The resources-issuing ROs must be backed by a physical resource that is capable of providing capacity when required but do not further constrain the nature of the capacity. Performance may be incentivized through both administrative penalties associated with failure to provide energy during peak hours, which may be set equal to or higher than the market prices, as well as the requirement that underperforming resources fund the procurement of alternative supply while missing out on scarcity rents.[63]

The advantages of the RO construct are similar to those of centralized capacity markets, including market-based pricing, transparency and resource neutrality. Notably, ROs also have the benefit of focusing incentives on only the highest price periods, and on energy market price signals, while providing long-term hedges to both suppliers and customers. The focus on price signals, rather than on placing behavioral obligations on producers (e.g., must-offer requirements), may lead some to view ROs as less administrative (i.e., less dependent on operator intervention and exogenous market rules) and more market-driven than capacity markets with physical obligations. Furthermore, limitations on energy market upside (particularly

62 See Peter Cramton et. al, “Capacity Market Fundamentals,” May 26, 2013, p. 8-9. To ensure the spot market energy and ancillary service prices fully reflect the scarcity conditions in real time, the Irish market is deploying an operating reserve demand curve in its reliability options market. The operating reserve demand curve is discussed in more detail in Section 3.2.

63 See supra note 51, at 34–35.

above the strike price) may serve to mitigate attempts to exercise market power in the energy market, and thus reduce the need for additional rules and monitoring efforts.

On the other hand, an RO market design has many of the same drawbacks as centralized capacity markets. Such a market is inherently administrative, and administrative decisions will affect its efficacy, including defining eligibility requirements, setting the appropriate procurement target, setting the strike price, and designing the auction. Among other things, there is concern that the strike price associated with the RO obligation will act as a cap on capacity resource offer prices as participants attempt to avoid penalty exposure should market prices exceed strike price. If this is true, and if the strike price is set too low, this could be a distortion of otherwise rational energy market offers, and incumbents will complain that the result is excessively low prices. Lastly, the financial nature of the obligation may be viewed as antithetical to those that view such capacity constructs as appropriately focused on ensuring physical reliability.

This approach is being implemented in Ireland, but as of yet, is untested.

3.2. "Energy-Only" Markets with Administered Scarcity Pricing

Capacity mechanisms are a means to address the resource adequacy problem. They are designed to compensate generators for the missing money in energy markets that arise from market and regulatory imperfections that suppress energy market prices during conditions of scarcity, and that arise from the suppression of scarcity by the pursuit of a physical reliability standard. The compensation comes from a second payment for capacity that is separate from, but in addition to, the payment for energy.

Hogan argues that while capacity mechanisms may address the resource adequacy problem (scarcity), by socializing the missing money through a common payment, they do not help with the core problems caused by inadequate real-time scarcity pricing.[64] Hogan recommends that a well-designed operating reserve demand curve (ORDC) would improve the problem of inadequate scarcity pricing. This would, in part, restore the missing money and improve capacity investment incentives. However, better scarcity pricing would have the added benefit of improving the incentives to operate the generation capacity efficiently, and change demand in response to short-term scarcity conditions.[65]

Hogan's approach builds on the existing use of real-time operating reserves. Operating reserve, which is generation capacity (or in some markets load reduction) that can be called on by the system operator in a short-interval of time (10 minutes to 30 minutes) to produce (reduce) energy in response to a sudden grid contingency, is a regular feature of electricity markets.

64 See William Hogan, "Electricity Scarcity Pricing Through Operating Reserves," *Economics of Energy & Environmental Policy*, 2013, p. 4.

65 The Electric Reliability Council of Texas (ERCOT) implemented an ORDC in June of 2014.

Similar to the previous discussion of resource adequacy standards and reserve margins, the system operator has an administrative requirement to hold a fixed amount of operating reserve in real time. The fixed level of reserves is equivalent to a vertical demand curve for operating reserve. The vertical representation of demand, however, does not reflect the actual economic value of reserves. The value of holding an additional MW of reserve beyond the fixed amount is positive and not zero as the vertical demand curve would suggest, as there is some reliability benefit from having the additional reserve to prevent potential loss of load. Similarly, falling just one MW short of the fixed amount cannot have an infinite value; a shortfall in operating reserves must have less value than the involuntary loss of load to which it is a contributing factor. And indeed, temporary shortfalls in operating reserves are expected outcomes of activating those reserves for the production of energy.

Hogan argues that if the markets were otherwise well functioning, with vigorous demand-side bidding in the energy market, the effect of this error in representation of the value of reserve in the dispatch algorithms would be small. However, the "chicken and egg problem" — the lack of adequate scarcity pricing inhibiting demand bidding — makes the error more important and requires a better representation of an ORDC.[66]

The key components of the design of the ORDC is the VOLL, the loss of load probability (LOLP) and the minimum contingency reserve requirement, The value of reserve anywhere along the ORDC is equal to the product of the VOLL and the LOLP. The minimum contingency reserve amount is the minimum amount of reserve that the system operator must maintain at all times to safeguard security constraints intended to prevent cascading failures that cause network collapses as described in Section 2.2.3. The system operator will shed load before running short of this minimum contingency reserve requirement.

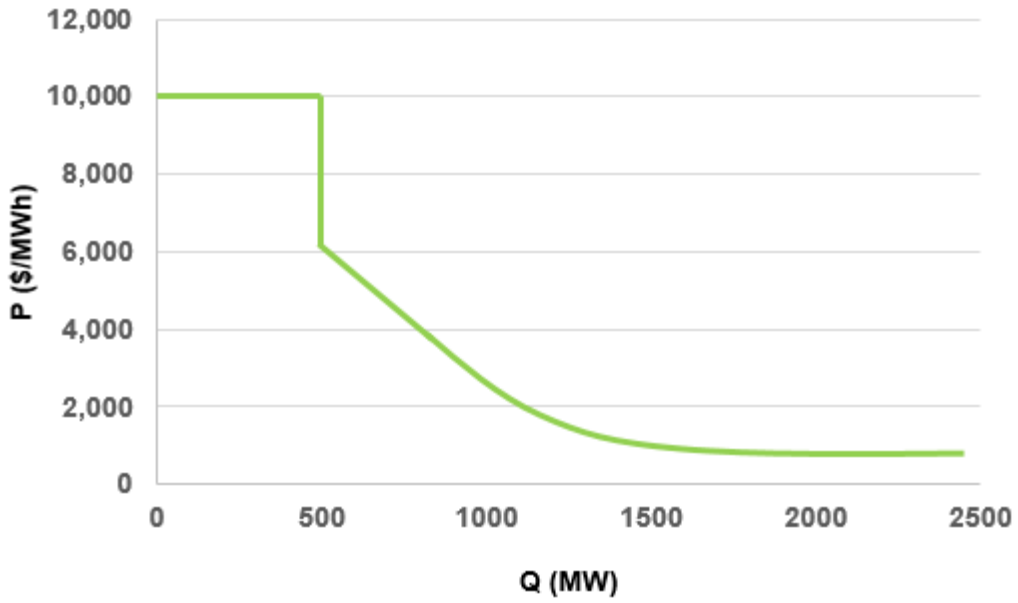
Figure 3 is an illustration of Hogan's proposed ORDC.[67] For this illustration, the VOLL is assumed to be $10,000/MWh and the mandatory minimum contingency reserve requirement is assumed to be 500 MW. As noted above, the system operator will shed load before running short of this minimum contingency reserve requirement. Therefore, for levels of reserve between zero and the minimum contingency amount (500 MW assumed in the illustration), the LOLP is administratively set to equal 1 and the ORDC is horizontal at the VOLL, reflecting that another MW of reserve would reduce a MW of involuntary load curtailment. For levels of reserve greater than the minimum contingency reserve amount of 500 MW, the ORDC reflects the incremental value of that reserve, which is the product of the VOLL and the LOLP. The more reserve that is available and procured, the lower the LOLP for the subsequent scheduling interval. Hence, the LOLP declines as more reserve is procured causing the value on the ORDC to decline. There is a discontinuity in the ORDC at the minimum contingency reserve of 500 MW. There is always a positive probability that load

66 *Id*., p. 70.

67 *Id*., p. 71.

could reduce in the next interval more than the expected generation losses, implying that the system operator will have more reserve available than the minimum contingency amount of 500 MW. This means that the actual LOLP at the 500 MW amount will be less than 1 when the system operator is operating at the minimum contingency amount. The discontinuity at the 500 MW amount is created by administratively imposing a LOLP equal to 1 and then using the actual LOLP (which is less than 1) for reserve amounts greater than 500 MW.[68]

Figure 3: Operating Reserve Demand Curve



The ORDC is implemented in the optimization algorithms and simultaneously optimized with energy. The simultaneous optimization means that the scarcity prices attributed to operating reserve apply to the energy price to reflect the trade-off between energy dispatch and carrying operating reserve. As a result, the ORDC produces energy scarcity prices at times of shortage and contributes to resolving the missing money problem.

There are many advantages to the ORDC.

68 For a detailed description of how the ORDC is implemented in ERCOT, see http://www.ercot.com/mktinfo/rtm/kd/Methodology%20for%20Implementing%20Operating%20Reserve%20Demand%20Curve%20.zip.

- The ORDC provides scarcity pricing for both energy and operating reserve that better reflects the immediate reliability conditions. This provides incentives for generation and DR to solve the actual reliability issues when they occur.
- The ORDC is fundamentally a "pay for performance" mechanism as it directly rewards performance for responding to the reliability conditions through the higher price of energy or reserve. This is something that capacity mechanisms strive to achieve but through less direct means.
- The ORDC affects the energy price and as a result, allows participants to hedge against increasing energy pricing. Hedging capacity payments is more difficult.
- The ORDC, by improving scarcity pricing and rewarding generation and DR directly for solving the reliability conditions, sends the right signals to invest in the types of operating characteristics that allow this response.
- The ORDC directly addresses the missing money problem and improves resource adequacy.

Similar to capacity markets, a disadvantage of the ORDC is the information and computational requirements for implementing the ORDC. A larger concern associated with the ORDC is that, unlike capacity markets that fix the quantity of capacity procured, there is no assurance that investment will occur to the levels needed to meet traditional reserve adequacy standards. Hogan argues, however, that the ORDC is compatible with forward-capacity markets and that there is no need to choose between the two. Addressing the fundamental issues related to inadequate scarcity pricing through the ORDC speaks to issues beyond the missing money, and can work in conjunction with a forward-capacity market if needed to achieve traditional standards.[69]

3.3. The Alberta "Energy-Only" Market

Another approach that relies on the energy-only market to drive investment is the approach that historically has been taken by Alberta. This approach allows generators to exercise market power unilaterally through economic withholding as a way to raise prices to levels that are expected to allow generators to cover their fixed operating costs and investment costs.

Arguably, this approach has worked for Alberta to date, in that it has attracted needed new investment and maintained a reliable supply of electricity.[70] That is, it would appear that Alberta has not suffered to the same degree from some of the market or regulatory imperfections discussed in Section 2.2 that have

69 The Irish Reliability Option model incorporates the ORDC within a capacity mechanism.

70 See https://www.aeso.ca/assets/Uploads/Albertas-Wholesale-Electricity-Market-Transition.pdf at p. 5.

caused the missing money problem in other jurisdictions. There may be several reasons for this.[71] For one, Alberta has not mandated a physical reliability standard such as the ones implemented in many other North American jurisdictions.[72] Instead, it has allowed market forces to establish the level of reliability. Second, and perhaps more importantly, unlike all U.S. jurisdictions, Alberta has allowed generators to recover their fixed costs and investment cost through economic withholding. Other influencing factors include political stability and positive supply, as well as demand fundamentals such as steadily growing demand, a relatively high load factor, and significant levels of cogeneration, all of which may diminish the effects of some of the shortcoming of energy-only markets described in Section 2.2.

There are pros and cons to this approach. On the plus side is simplicity — there is no need for complex administrative measures such as capacity mechanisms or operating reserve demand curves. The disadvantage is that resource adequacy is determined by the market, rather than assured through mandated standards. This may eventually lead to periods in which there are politically unacceptable levels of load shedding. Furthermore, periods of prolonged scarcity could lead to the extreme exercise of market power and associated loss of allocative and productive efficiency that may or may not be offset by future dynamic efficiency gains from new entry.[73]

The CLP will bring about significant changes to the future supply mix in Alberta. The pending changes introduce uncertainty of future revenue streams and raise the risk associated with new project financing. The stability of the energy-only market with respect to reliability and prices under the new directed supply mix is unknown. It would appear that these factors may have influenced the government's decision to implement a capacity market to ensure stability with respect to reliability and prices.[74]

---

71 See Brian Rivard and Adonis Yatchew "Can the Electricity Market Structure Accommodate Significant Levels of Renewable Generation? An Evaluation of Carbon Policy Options for the Alberta Electricity Sector," Paper prepared for the Alberta Market Surveillance Administrator, October 2015, for a discussion of possible reasons. *See also* Matt Ayres. "Making 'Energy-Only' Markets Work," Harvard Electricity Policy Group 71st Plenary Session, June 13, 2013. Calgary, Canada.

72 See https://www.aeso.ca/assets/downloads/Long-Term-Adequacy-Feb-7.pdf, p. 4, where the AESO notes that "In an energy-only market design, the market determines the appropriate level of adequacy over the long term." See also Hannes Pfeifenberger, "Resource Adequacy Requirements, Scarcity Pricing, and Electricity Market Design Implications," presented to IEA Electricity Security Advisory Panel (ESAP), July 2, 2014, slide 4, available at https://www.iea.org/media/workshops/2014/esapworskhopi/Hannes_Pfeifenberger.pdf.

73 In Alberta, the Market Surveillance Administrator is responsible for the monitoring of extreme instances of market power, with the authority to investigate and possibly take enforcement action against such conduct.

74 As stated at on the Alberta Government's website at https://www.alberta.ca/electricity-capacity-market.aspx, "A "capacity market" for electricity will protect consumers from price volatility and provide a reliable supply of electricity at stable, affordable prices."

3.4. Summary of Alternatives to Addressing Missing Money

In summary, there is no universally agreed-upon approach amongst scholars and practitioners on how to address the resource adequacy or missing money problems. Each has its pros and cons. That being said, most jurisdictions have chosen or are choosing to implement centralized capacity markets. Furthermore, as Ireland has exhibited, combining both a centralized capacity market — in its case, the reliability options model — with the ORDC is also possible. Such an approach addresses the resource adequacy problem while also improving the market-based incentives to operate the generation capacity efficiently and change demand in response to short-term scarcity conditions. In this manner, it is desirable to supplement administrative mechanisms with continued efforts to remove as many imperfections from the primary market as possible, as doing so promises to limit dependence on the secondary market, thereby limiting continued litigation regarding its parameters.

4. The Economics of Basic Capacity Market Design Features

This section will focus on design elements of centralized capacity markets. No two capacity markets are the same; all have evolved in the context of pre-existing institutions and regulatory frameworks. In this regard, there is no single accepted theoretical view of the optimal design of a capacity market. Rather, both the motivating factors and the implementation details of capacity markets vary by jurisdiction. However, the economic and academic literature provides some guidance on the choice of basic design features and provides insight as to the discrete trade-offs associated with market rules. In this section, we review the economic principles and implications behind these basic design features.

4.1. Capacity Product Design

A capacity market requires a definition of the product to be exchanged between buyers (or in the case of some centralized capacity markets, a single buyer) and sellers. In the context of a capacity market, there are certain attributes generally considered when defining the product. First, the product should be defined in a way that is consistent with the reliability objective for which it was created, which is resource adequacy. Most jurisdictions have defined the capacity product generically as the "availability to generate energy or reduce load when needed," typically during periods of shortage or scarcity.[75] This is generally expressed as Unforced Capacity (UCAP), a measure of the maximum ability to generate, adjusted for a factor representing a resource's historical availability called the Equivalent Forced Outage Rate (EFORd). However, the UCAP metric has faced criticism, with opponents arguing that EFORd measures availability

75 As discussed below, while the definition focuses on the availability of capacity, what is actually needed is the actual energy from that capacity during the periods of shortage or scarcity. One of the challenges that many capacity markets have faced is providing the appropriate incentives to supply-side and demand-side resources to actually perform (produce energy) during these periods. See discussion on performance obligations in Section 4.6.

too generally, and a better metric would be more focused on availability at times when the system is actually short of operating reserves or energy. An alternative would be to rely on total installed capacity (at least for traditional resources) as the product procured, complemented with strong incentives and penalties associated with actual performance (as discussed in Section 4.6).

Second, in order to promote economic efficiency and competition, the capacity product should be defined in a way that is neutral to the technology used. Such neutrality is achieved if the capacity product is defined in such a way that a MW from each resource represents an equivalent reliability value. Rules that achieve this objective ensure that the ultimate resource mix is driven by economics rather than by administrative fiat. To the extent that it becomes apparent that market rules — often in the form of resource qualification requirements[76] — are dictating outcomes and not allowing deserving resources to participate on an even playing field, those rules should be changed, if at all possible, to foster competition. Of course, not all resources are equivalent. Variable energy resources (e.g., wind and solar) and DR, for instance, have characteristics that complicate direct comparison with traditional thermal generators. Nonetheless, market rules should ensure that each resource can qualify to offer and if cleared, be compensated for a quantity of capacity that reflects its ability to contribute to system reliability.

Third, in some jurisdictions, the capacity product may require a locational attribute if transmission constraints limit product deliverability from one location to another. However, such differentiation is focused on recognizing the value of capacity in relation to its location within the system topology; the product definition itself does not vary geographically. Locational pricing is discussed in detail in Section 4.3.

As system needs have changed to reflect the additional demands resulting from high penetrations of variable resources, an emerging issue is whether the basic definition of the capacity product should be changed, or be split into multiple discrete products to ensure that system needs are fully provided for. For example, should the capacity market explicitly procure resources that are able to ramp quickly, start on short notice, or provide fast regulation services (i.e., the ability of generators to modulate output very quickly — on the order of seconds — particularly in response to varying output from renewable energy resources)? Alternatively, should capacity markets explicitly seek resource diversity and limit how much capacity can be procured based on one fuel type? FERC staff point out, “redefining the capacity product to procure needed operational attributes, defining the specific attributes to be procured, and determining how much of the overall capacity requirement should be met by capacity resources with such attributes would be a

76 A resource qualification requirement is a requirement that a resource must meet in order to qualify as a capacity resource. For example, a qualification requirement that a capacity resource have a contracted fuel supply might inappropriately exclude the ability of renewable or demand-side resources to participate in the market.

complex undertaking."[77] Indeed, accomplishing these additional and varied objectives would add significant complexity to be managed by buyers, sellers, and market operators in capacity markets. More broadly though, subdividing the capacity product promises to increase administrative intervention into an already heavily administrative market, thus increasing the likelihood of inefficiencies. Rather, efforts to stimulate investment in resources with attributes that help optimize system operation are best directed at improving the accuracy of prices in the energy and ancillary service markets, as described further in Section 4.10.

4.2. Auction Format

There are two common formats for capacity market auctions. Additionally, hybrid structures are possible that allow realization of beneficial features from both primary formats:

- **Sealed bid auction (SBA):** In an SBA, resource owners submit sealed bids at the closing date and time. Offers are final and irrevocable and specify, at a minimum, the quantity of capacity available at what price. These are assembled into a supply curve, which the auction process then matches to the demand curve to establish a clearing price. All offers below that clearing price are accepted and the associated offerors take on capacity obligations.
- **Descending clock auction (DCA)**: A DCA format is more interactive. The auctioneer posts a starting price (generally the price cap level based on the demand curve). Participants specify the capacity resources they would provide at that price. The auction then proceeds in successive rounds. In any given round, if the total capacity offer quantity exceeds the capacity demand at that price, the auctioneer repeats the process at a lower price, and participants update their capacity offer quantities at the reduced price. The process continues until the total capacity offer quantity drops below the demand curve quantity at that price, thus establishing a clearing price and the identity of resources cleared.

Each of these approaches is in use in North American power markets and internationally. Examples of SBA markets include PJM, NYISO and Ireland, while ISO-NE and GB utilize a DCA approach to secure commitments.

There is no consensus as to which auction format is superior. There are pros and cons to each.

An important element in capacity market design is mitigation of market power; ensuring that the larger suppliers cannot influence prices to their benefit and to the detriment of the overall efficient operation of the market. Some economists argue that the DCA auction is more susceptible to the exercise of market power as the succession of rounds allows bidders to learn when its closest competitors exit the auction and when its bid may become pivotal and set the price. When this is the case, the bidder may adjust its strategy to

77 FERC Staff Report. "Centralized Capacity Market Design Elements," 23 August, 2013, p. 18, available at https://www.ferc.gov/CalendarFiles/20130826142258-Staff%20Paper.pdf.

raise its bid price above its true cost. Additionally, bidders may also learn through the succession of rounds that the supply situation is tighter than they had expected prior to the start of the auction. This can result in bidders adjusting their strategy, potentially setting prices above their true cost. [78]

In an SBA, the incentive for participants to offer at prices that best reflect actual cost serves as a mitigation tool for market power. SBAs maintain higher levels of confidentiality, as they require only a single instance when an offer is placed, and do not employ successive rounds in which participants can observe the bids and strategies of their peers and modify their own. In this way, SBAs are less likely than DCAs to clear at prices higher than generation costs, because participants are less likely to attempt to influence the auction outcomes by adjusting their bids based on observed behavior by other participants. This also mitigates risk for market participants, because if they clear the auction, they are assured to receive at least the level of revenue specified in their offer. If they do not clear, they are saved from having to operate at prices that would be expected to lead to a loss (i.e., prices below their offer price).

On the other hand, the DCA format may be preferred if the auction participants are uncertain about the future cost of delivering the goods sold.[79] Given the volatility of commodity prices, it is very possible that some bidders incorrectly calculate their expected future generation costs. By observing other comparable participants' actions in the market, a bidder can infer what the average expected costs are, and adjust its strategy accordingly.[80] This format aids both over- and under-estimators: a bidder with higher cost expectations than the rest of the field can observe excess supply after a round and lower its offer price; under-estimators can observe competitors exiting the market earlier than they had planned and adjust their bids. On the other hand, in pure SBA formats, the over-estimators may be left out of the market, and under-estimators, while cleared and in the market, may struggle to recover their costs later on if their bids result in a lower market price.

A final consideration for the auction design is its simplicity. The format of the SBA is much simpler as it requires less time and fewer resources to implement.

Given the pros and cons of each type of auctions, some market operators employ a hybrid of the two. For example, ISO-NE uses a hybrid DCA format, which includes bid reviews prior to the auction and real-time adjustments allowed by the auctioneer. The hybrid DCA format continues to give market participants insight into excess supply after each round, protecting from mispricing errors in offers that can ultimately lead to

78 See ISO New England. "Forward Capacity Auction Formats," July 2016, p. 5, available at https://www.iso-ne.com/static-assets/documents/2016/07/20160711-dca-v-sealed-bid.pdf.

79 *Id*., p. 5.

80 Whether such strategies can be executed in practice is debatable. Among other reasons, given time constraints of the auction and the mix of participating technologies, it may not be feasible to actually infer significant information from the actions of other auction participants.

inefficient market outcomes. However, it better mitigates market power exercise by limiting the bidding latitude on existing resources (they must bid zero), and by allowing new resources to respond to the new information at each successive round.[81] With this, the auction is less susceptible to price manipulation but preserves the cost risk mitigation for new resources bidding into the market. The hybrid auction format contains sealed bids within each round, so that participants cannot see what others have entered. Like a pure SBA, this incentivizes participants to bid at their true costs, reducing the likelihood of clearing prices above cost and alleviating some concerns over exercise of market power.

As described, cleared capacity market offers by participating resources are generally paid a single clearing price, regardless of the type of resource that supports the ability to sell capacity. This outcome is supported by both economic theory and the motivating purpose of capacity markets. First, like energy spot markets (and all other commodity markets), use of a single clearing price for all market participants is not only simple, but the efficient means of determining which producers should produce and the associated price — that is, how much they should be paid.[82] The same theory that supports this approach in energy market pricing applies to the conduct of auctions in capacity markets. Second, the question often arises as to whether it is appropriate for all types of capacity resource — as a logical matter — to receive the same price, regardless of whether a unit provides baseload-type power, peaking-type power, or otherwise. The answer lies in whether all units in the market are equally affected by shortcomings in the energy markets that affect energy market outcomes, and lead to missing money. Indeed, they all are. Thus, to the extent that the missing money problem is experienced uniformly by all resources, the same capacity payment should be received by all resources to remedy this problem. In this manner, the ultimate stream of payments — energy, ancillary services and capacity — received by all resources in the market (with capacity obligations) should add up to the level of payment that approximates what would be available in a perfectly competitive market.

4.3. Zonal vs. Uniform Pricing

Zonal capacity pricing is implemented by separating the broader system into capacity zones based on the location of existing capacity and the ability to transfer capacity between geographic areas. This includes defining specific locational capacity requirements and inter-zonal transmission limits. A locational approach is consistent with energy markets that account for geographic variation in price based on generator locations and transmission constraints, although capacity zones are typically less granular than any energy zones.

---

81 See *supra* note 78, at 7; see also *supra* note 35, at 22.

82 This result is born out extensively in the economic literature. For an energy-specific reference, see, for example, Peter Crampton, “Single Clearing Price in Electricity Markets,” February 18, 2009,available at ftp://www.cramton.umd.edu/papers2005-2009/baldick-single-price-auction.pdf.

In a zonal capacity market, demand curves are established for each area as well as the total market; locational demand curves may recognize differences that exist in net CONE due to siting and construction costs, or different energy price expectations. In this type of market, auction-clearing prices are uniform among zones, between which, transmission limits are not binding; zonal prices separate where transmission limits constrain deliverability of otherwise competitive resources. Zones relying on imports may set minimum MW requirements that must be served by resources within the zone. If sufficient capacity is offered, the import constraint binds and the price in that zone is fixed. In export-constrained zones, if export constraint binds, then the price falls until the constraint is resolved.

Markets with zonal pricing better allocate costs and incentives for investment, allowing better alignment between missing peak-hour energy revenues and the associated capacity market price signal. They address the need to balance the increased competition possible through system-wide mechanisms with the locational requirements of reliability. If the system-wide clearing price does not attract sufficient resources to meet the requirements for a zone (inclusive of transmission imports), a locational price adder can better reflect the value of local resources. Clearing congested zones separately, and at higher prices, increases the likelihood that adequate resources are incentivized to offer where needed. In particular, higher capacity payments may be necessary to promote new development in constrained regions, where additional new generation is often required but also more difficult to build. Use of zonal pricing in capacity market design may create the opportunity to stimulate efficient investment choices in transmission; some capacity markets allow transmission developers that increase capacity to constrained areas to earn capacity payments for their investments.

The alternative to locational pricing is uniform capacity pricing. While uniform pricing may increase equity among all generators in a market, it can lead to distortion of prices. It can also lead to a failure to send locational signals that promote development where needed and that reflect the relative value of existing resources. However, a single zone increases simplicity in the grid and auction, as well as resource planning. Uniform pricing may be more appropriate in systems where there is an abundance of transmission capacity and no transmission constraints between areas within the system. Alternatively, in a power market that does not provide locational signals for energy, it may be difficult to justify and/or implement a capacity market that differentiates prices by location.

All U.S. jurisdictions with capacity markets operate locational or zonal capacity markets in which the capacity price can vary between zones. This reflects the locational pricing structure of US energy markets. The zonal capacity markets recognize transmission constraints that may limit the ability of energy to be delivered between zones under stress conditions. In contrast, Great Britain and Ireland operate uniform capacity markets consistent with their energy markets, although Ireland recognizes deliverability limitations in procuring capacity and makes “out of merit” payments to attract the required capacity in the constrained zones.

Alberta currently operates a uniform energy market. It also has a "congestion-free" transmission policy, whereby the AESO is required to ensure that the transmission system internal to Alberta is appropriately reinforced, so that under normal operating conditions, about 95% of expected wholesale transactions can be realized without transmission congestion.

The broader transmission policy that supports the uniform energy price does have other interactions with the capacity market design. On the one hand, current policy does not pose immediate issues for the capacity market to the extent that it has led to no material transmission constraints that limit deliverability of existing generation in times of system stress. However, with significant amounts of retirement and new entry pending, the "congestion-free" policy may not send the appropriate signals to incentivize the socially efficient location for new generation to build in the future.

4.4. Forward and Commitment Periods

The capacity market forward period is the period between the completion of an auction and the start of the capacity delivery period, which is also referred to as the commitment period. The commitment period is the length of time for which a seller that clears the auction is required to meet its obligations and provide the capacity product to either the system operator or the LSE. The selection of the forward period and the commitment period can affect the efficacy of the capacity market toward ensuring reliability, and it can also affect overall market structure, financing decisions, and the type of resources that are procured.

Forward periods can range from one month to several years. Of course, estimates of required capacity and required reserve margins are more accurate as the delivery period nears. Thus, shorter forward periods will result in a closer match between the capacity auction results and the quantity of capacity ultimately needed. More precision in procurement serves to mitigate over-procurement risk for customers, as well as for demand-side resources, which may face difficulties in securing commitments from individual customers too far in advance of the delivery period. However, particularly when significant new capacity is needed, a capacity market with a short forward period may face shortfalls if there is insufficient time to complete construction of new power plants ahead of delivery, let alone to account for delays that may arise and push back online dates.[83] The relationship between the forward period and the time it takes to develop different types of resources may also drive results; the shorter the forward period, the more the market will skew towards short lead-time resources. Longer lead-time resources will often need to enter bilateral contracts (outside the capacity market) spanning multiple delivery years in order to achieve an acceptable level of cost-recovery certainty.

83 See FERC Staff Report. "Centralized Capacity Market Design Elements," 23 August, 2013, pp. 12–13,available at https://www.ferc.gov/CalendarFiles/20130826142258-Staff%20Paper.pdf.

Longer forward periods provide more adequate lead-time for capacity resources to be constructed and come online, and better determine supply of new builds. In turn, this may encourage new entry to the market, as resources at various stages of planning and construction can offer their capacity and secure commitments. A long forward period provides early-stage developers increased confidence that they will recover their costs, and provides time for existing resources to decide whether to exit the market or to retrofit a plant.[84] The associated reduction in risk may also reduce the cost of capital. Unfortunately, longer forward periods also leave more time for unexpected shifts in the forecast reserve requirement, particularly as demand fluctuates. To correct for this, capacity markets with long forward periods often conduct "true-up" auctions between the initial auction and the beginning of the delivery period, or allow participants to trade or sell commitments bilaterally. They may also have rules to procure most, but not all, of the expected capacity requirement in the initial auction, and then employ supplemental auctions closer to the delivery date to procure the remainder.

The commitment period may be annual, seasonal (summer and winter) or monthly. Some jurisdictions also offer multi-year commitment periods to new build generation. Similar to the forward period, the selection of the commitment period can also affect how capacity suppliers engage in the market, especially new resources. Longer commitment periods provide more certainty and incentives to shoulder the risk of building new generation, as it is more likely that developers will expect to be able to recover costs. Longer commitment periods also mitigate short-term volatility by setting prices for extended periods of time, but they also reduce the frequency and precision with which price discovery and supply-demand balancing takes place. Market regulators generally attempt to strike a balance between providing additional certainty in cost recovery and maintaining balance in the marketplace.[85] In order to account for the financing needs of new resources, some capacity markets include provisions for extended payment guarantees should a new resource clear the market and opt to lock in the price for more than one delivery period.

There is less often emphasis on long forward and commitment periods in markets with decentralized capacity obligations imposed on LSEs. This may be explained by the fact that, in those markets, the LSEs provide the bridge between shorter capacity market terms and longer investor needs, confident in their ability to recover prudently incurred costs under their state regulatory regimes.

Design decisions and regulatory actions that affect real or perceived capacity market volatility may also affect the type of investment decisions that are made. A recent ISO/RTO Council (IRC) study, based largely on interviews of personnel involved in investment evaluation and decision-making, revealed that decisions are largely driven by energy revenues and the belief that energy demand will grow or that supply will

84 *Id.*

85 *Id*., p. 14.

decrease with plant retirements due to age or environmental regulations.[86] In jurisdictions that operate capacity markets, investors have expressed their support of capacity markets as they provide a backstop to energy revenues and greater certainty of future cash flows. However, investors have raised a concern regarding the volatility of capacity market prices and hence revenues, and the uncertainty that this creates. The IRC study notes that capacity market price volatility has caused many financial institutions to discount capacity revenues heavily when considering financing new investment projects. By comparison, "Those with a longer-term investment outlook, such as balance sheet-backed investors, indicated that they discounted projected capacity revenues less heavily, and also felt that they had a better capability to model these revenues." [87]

4.5. Capacity Cost Allocation

There are several levels of cost allocation and settlement associated with capacity charges. At the wholesale level, charges for capacity procured are allocated to LSEs (or other retail suppliers) and other wholesale market participants based on market rules. It is these payments that are ultimately distributed to capacity providers at the settled price. Moreover, it is this payment stream that is likely to be subject to any performance adjustments or deficiency charges based on what is actually delivered during the commitment period. Capacity costs borne by LSEs and other wholesale capacity customers may then be passed on to retail customers through retail rates and tariffs subject to the relevant regulatory body. In addition to the primary settlement processes, capacity market participants may undertake bilateral transactions, which may take the form of private commercial transactions for capacity or of hedging arrangements against auction clearing prices.

For the purposes of capacity market design, the most relevant aspect of cost allocation is that of wholesale market charges to capacity market buyers as this is generally what is regulated by the entity that approves the market rules. Capacity charges may be allocated based on peak energy demand and overall energy consumption, or may be bundled with other system charges before being assigned to market participants.[88] Generally, capacity needs are viewed as being driven by the need to serve peak load, which suggests that an efficient allocation to cost causers should be based on peak load requirements. Jurisdictions such as PJM use a five coincident peak allocation method, whereby total capacity costs are allocated across

86 A recent report commissioned for the ISO/RTO Council concluded that the key driver of investment in any market is strong supply and demand fundamentals. See http://www.isorto.org/Documents/Report/201505_IRCResourceInvestmentReport.pdf at pp. 18–19.

87 *Ibid.* It should be noted that the investment risks of price volatility in an energy-only market may be more acute than the investment risk from capacity price volatility, a fact that provides insight into why there was support for capacity markets by investors who were interviewed as part of the IRC study.

88 See supra note 51.

individual customers based on their share of energy consumed during the five coincident peak demand hours during a defined annual recovery period. An allocation based on peak load arguably provides efficient price signals to customers regarding the system and the reliability effect of their consumption patterns.

That being said, some have noted that cost allocation rules should account for the relationship between customer demands and the need to have capacity, and that the coincident peak methodology may miss important features of this relationship.[89] For example, the need to have capacity may be evident in hours that are not necessarily the system peak-demand hour, such as hours when, due to outages, there is limited supply for the prevailing demand. In these hours, the energy price may be very high, reflecting the relative supply shortage. In an energy-only market, where capacity costs are recovered strictly through the energy price, this is the signal that new capacity may be required. Arguably, a cost allocation method that aligns with the energy-only market and recovers costs during these hours, may provide an efficient price signal to customers regarding the system and reliability effect of their consumption patterns.

Policy decisions, however, may dictate other approaches to allocating capacity costs. For one, capacity costs may be associated with total energy consumption rather than contribution to system peak. This would result in an allocation that shifts costs from wholesale buyers with lower load factors (e.g., load serving entities serving predominantly residential load) to those with higher load factors (e.g., industrial customers participating at the wholesale level). This method dampens capacity signals associated with peak management, and instead, ties capacity cost management to reductions in overall consumption. Alternatively, capacity costs may be lumped in with other system charges (e.g., market operator fees, transmission fees, etc.) and then allocated in whatever fashion has been approved for the overall set of system charges. While this may be administratively simple, such an approach both fails to convey efficient economic signals about the cost of capacity, and reduces transparency as to the cost of capacity being funded by consumers. Moreover, allocating capacity costs in this manner could allow loads with behind-the-meter generation to avoid other system charges, including capacity charges, which would then be placed on other customers.

Once capacity costs are allocated, they may then be charged in one of two ways to wholesale capacity buyers during settlement. First, allocated costs may be charged on a per-MW basis. A single charge for capacity — often on a monthly basis — is efficient (at least relatively, depending on the allocation method), transparent and administratively expedient. Wholesale customers see, at one time and in one payment, how their decisions drive capacity costs. Second, capacity may be wrapped into energy prices and charged on a bundled, per-MWh basis. This has the benefit of conveying to system customers the all-in cost of

89 See "Design Considerations for an Alberta Capacity Market" a paper prepared by Monitoring Analytics for the Alberta Surveillance Administrator, September 21, 2016, p. 11, available at http://albertamsa.ca/uploads/pdf/Archive/00000-2017/2017-01-18%20Alberta%20Capacity%20Market%20Report%2009.21.16.pdf.

providing reliable wholesale electricity service. However, this method of charging fails to provide transparency as to the separate price of procuring capacity. It leads to an inefficient cost allocation, as not all MWh consumed are equally responsible for driving capacity needs, despite what would be implied to consumers by an all-in price. Finally, a consumption-based capacity charge adds administrative complexity; it requires forecast of expected load to determine a per unit capacity adder, and if the forecast is wrong cross-period, true-ups will be required to maintain a balance of payments.

4.6. Performance Incentives and Obligations

Capacity markets generally provide a uniform price for the capacity product itself across the commitment period. However, long-term price signals such as capacity payments do not provide short-term incentives to perform when called upon during periods of shortage or scarcity, nor do they promote investments to ensure availability during such times. After all, committed capacity has limited value if resources with capacity commitments cannot, or will not, perform during periods of system stress. To some extent, higher prices during periods of scarcity will convey signals to induce good behavior. However, in a system with a capacity market, peak energy and ancillary service prices are lower than in an energy-only market, thus suppressing signals for efficient investments and activities.[90]

To ensure that resources are providing capacity value when it is most needed, most capacity market constructs are developed with incentives and obligations associated with availability and market offers during peak periods. These may include:

- **"Must-offer" requirements:** To ensure that capacity resources are available to provide energy or ancillary services when needed by the system, capacity resources are often required to offer into the energy and ancillary service markets when they have not notified the system operator that they are otherwise unavailable for an allowed reason. (A must-offer requirement is also a tool to prevent physical withholding and attempts to exercise market power.)
- **Long-term incentives:** Units that perform at their peak capability frequently throughout the year, or particularly during peak periods, may be allowed to sell additional capacity in future delivery periods, while units that underperform may see their future sales limited. Often, this is implemented through the calculation of equivalent forced outage rates (EFORd), which in turn, are used to calculate unforced capacity (UCAP), which may be sold into the capacity auction.
- **Short-term incentives:** During tight conditions, rules may be developed such that supply resources face additional opportunities and risks associated with performance. A rule set like this will often define

90 See Peter Cramton and Steven Stoft, "A Capacity Market That Makes Sense," *Electricity Journal*, 18, 43–54, August/September 2005, p. 6.

particular, condition-driven periods during which additional, often severe, penalties apply to resources that underperform. Likewise, units that over perform may have an opportunity for bonus payments.

An effective system will not only promote good behavior during periods of system stress — encouraging availability and efficient market offers — but will also incent resources to make efficient capital investments and business decisions, ensuring they are able to operate reliably when needed most. Such measures might include upgrades to increase reliability and decrease unscheduled outages, improve forward planning and coordination to avoid planned outages during likely performance assessment periods, and fuel arrangements to ensure supplies are available even during disruptive events.

4.7. Market Power Mitigation

4.7.1. Seller-Side Market Power

Supply-side market power arises from concern over scarcity of supply, leading to the ability and incentive of certain suppliers to bid above cost. This, in turn, leads to capacity prices above efficient levels. This concern is particularly acute in regions that are delivery constrained, a characteristic that is often paired with local challenges to developing new generation. There are two capacity design elements that are employed to manage supply-side market power: 1) active mitigation and 2) demand-curve design. Additionally, as described in Section 4.2, the selected auction format may play a role in addressing concerns over exercise of market power by capacity suppliers. When addressing seller-side market power, care is necessary to ensure that mitigation mechanisms do not suppress prices below competitive levels, thus risking the ability of the market to attract and retain appropriate levels of capacity.

Generally, active mitigation has two steps. First, the market operator runs a test to determine if a particular capacity resource has the potential to exercise market power or, alternatively, if the market as a whole (or in a constrained region) is susceptible to the exercise of market power. These tests will often look for pivotal suppliers[91] — individually or in groups — or instances in which certain suppliers have particularly high market shares. Second, when tests indicate that market power might be a concern, offers by any suppliers of concern are mitigated. To what level offers are mitigated, a controversial topic that can account for a number of scenarios (e.g., retirement, mothball and failure to clear but continued operation), is usually based on unit-specific determinations of going-forward costs derived from additional information submitted by market participants. After the market operator or market monitor informs the market participant of its mitigated bid level, the participant may be allowed to challenge that level and ask for evidence-based

91 A supplier is pivotal if its output is required to meet demand. PJM uses a three pivotal supplier test to assess the market power in a relevant market. The three pivotal supplier test measures the degree to which the supply from three suppliers is required in order to meet the demand in the relevant market.

exemptions from formulaic mitigation levels. Whatever mitigated price is determined to be appropriate for mitigated capacity resources, that price represents the resources offer in the capacity market.[92]

The design of the demand curve can also contribute to efforts to mitigate market power. Specifically, selection of a sloped demand curve that is elastic around the target capacity level serves to limit the ability of individual market participants to dramatically affect the auction-clearing prices. Sloped demand curves limit how large a price increase can result from small changes in available supply. As a counterexample, selection of a vertical demand curve can lead to extreme price volatility around the reliability target, which exacerbates the ability of pivotal suppliers to influence significant changes in market outcomes. Design of supply curves with maximum (and minimum) prices also assists market operators in constraining prices to levels that are determined to be economically appropriate given system conditions.[93]

Seller-side manipulation may also be addressed, in part, by constraining which resources are allowed to offer into the market. For example, ISO-NE has implemented a system in which existing capacity resources that intend to continue to operate, regardless of capacity market price, are simply not allowed to submit bids in the annual auction. Such resources are assumed to be price-takers; eliminating them from active participation reduces the number of market participants that could attempt to exercise market power. Crampton and Ockenfels also suggest the possibility that all generators be required to offer — a "must-offer" requirement — unless they provide justification (e.g., unreliability or sales to other systems) for their actions. This approach mitigates concerns over withholding, but may be cumbersome to implement, particularly as new resource types become common in the supply mix and face challenges over subjectivity.[94]

4.7.2. Buyer-Side Market Power

Buyer-side market power is a concern when market participants that are net-buyers have the incentive and ability to suppress prices below efficient levels. Such an outcome can be achieved by a net-short market participant offering its supply capacity (if a vertically integrated player) at a price below its actual cost, thus potentially driving down market prices and reducing the overall cost of its capacity market position. As with supply-side market power, constraining the ability of net-short market participants to exercise market power is important to maintaining competitive market outcomes. Likewise, buyer-side mitigation rules first identify

92 See "Centralized Capacity Market Design Elements," Federal Energy Regulatory Commission Staff Report, Docket AD13-7, August 23, 2013, pp. 22–24.

93 See Peter Cramton and Axel Ockenfels, "Economics and Design of Capacity Markets for the Power Sector," 2011, pp. 20–21.

94 *Id*., p. 19.

resources of concern, and then mitigate their bids *upward* to levels deemed competitive, all of which are usually accomplished by some form of Minimum Offer Pricing Rule (MOPR).

Under a MOPR, capacity resources of concern are not usually identified based on structural metrics, like whether they are pivotal or have high market shares, but by specific characteristics of the resource that indicate whether it is well suited to exercising buyer-side market power. A non-exhaustive list includes:

- New resources, which can be offered below cost and shift the existing supply curve to the right
- Resource types that are well suited to exercise of capacity market power (e.g., gas combustion turbines and combined cycle generators), largely due to low development costs and short lead times
- Resources in particularly tight capacity zones where new low-cost capacity is more likely to affect major (downward) swings in price
- Resources receiving out-of-market subsidies

Resources that are selected for mitigation have their offers reviewed, and if necessary, adjusted to ensure that offer prices are in line with what may be considered competitive levels. Usually, this takes the form of a bid floor, thus the *minimum offer* aspect of MOPR. Mitigated prices are usually predetermined and based on the expected net CONE, which may vary by resource type. Market rules may also allow for exceptions to resources that would otherwise be mitigated; for instance, if an offer by a specific resource is demonstrated to be competitive or for self-supply.

While important, MOPRs have proven to be difficult to establish and implement in a way that remains robust to policy interventions. Numerous instances have arisen in which MOPR rules have run afoul of national or regional energy policies (or vice versa), like generation subsidies or renewable portfolio standards. When resource procurement is affected by economic or regulatory influences external to wholesale energy markets, the associated capacity resources can end up being subject to mitigation. Mitigated resources may not ultimately clear the market, which will in turn procure sufficient capacity from other sources. The risk is that such an outcome results in over-procurement, as the capacity market procures levels of supply deemed appropriate to provide sufficient resource adequacy, while separate resources are *also* procured and paid for to fulfill public policy requirements. The same concern arises if an LSE opts to self-supply new capacity when market rules do not provide exceptions provided for such behavior. Of course, market rules can be designed to account for certain circumstances,[95] but additional exceptions reduce transparency, lead to regulatory uncertainty, and prompt complaints over discrimination. As it is unlikely that energy policy

95 For example, resources receiving subsidies could be excluded from the capacity market, while also reducing the target reserve margin by a commensurate amount. Likewise, a self-supplying vertically integrated utility could have only its net needs reflected in the capacity market. There are a number of approaches to addressing subsidized generation and self-supply, with the issue of subsidized generation being particularly relevant at this time in North American markets.

interventions subside, the tension between accommodating non-market energy policy and fostering efficient achievement of resource adequacy through capacity markets, is likely to persist.[96] How capacity market design can coexist with exogenous public policy goals is a topic of ongoing discussion that has yet to yield any best practices. This is discussed further in Section 4.8.

4.7.3. Energy Market Offer Mitigation

While not officially a feature of capacity market design, it is worth noting that all U.S. markets that operate capacity markets employ market power mitigation measures in their energy markets.[97] These mitigation measures, by design, can affect energy market price levels, and consequently, will influence the level of revenue generators can receive in the energy market. This, in turn, affects the level of revenue that the generators need to recover in the capacity market and hence the generators' bids in the capacity market.

The energy offer mitigation measures are designed to ensure that resources are able to offer energy at their marginal cost but are not able to exercise market power. The details of the market mitigation measures differ in each market. However, there are two general approaches: (1) structural, and (2) conduct and impact.[98] PJM and the California Independent System Operator (CAISO) use the structural approach, while ISO-NE, NYISO and MISO use the conduct and impact approach.

Under the structural approach, resources are subject to offer mitigation when predefined conditions are met in the market as a whole. For example, PJM uses a three pivotal supplier test that assesses whether the three largest suppliers in a relevant constrained region of the market are jointly necessary to relieve the constraint. If so, the energy offers of these suppliers are mitigated to equal a reference level offer,[99] which is an estimate of their marginal cost plus a 10% adder.

The conduct and impact mitigation approach involves two steps. The first step is to establish whether a resource's offer exceeds its reference level offer (an estimate of the resource's marginal cost) by a pre-specified amount. The second step is to assess the impact of that offer price on the market-clearing price. If the resource's offer fails the conduct test, and the impact of its offer exceeds a pre-specific impact threshold, that resource's offer is replaced with the reference-level offer in the operation of the energy markets algorithms.

---

96 See "Centralized Capacity Market Design Elements," Federal Energy Regulatory Commission Staff Report, Docket AD13-7, August 23, 2013, pp. 24–28.

97 For a summary of approaches, see https://www.ferc.gov/legal/staff-reports/2014/AD14-14-mitigation-rto-iso-markets.pdf.

98 All markets discussed here (CAISO, ISO-NE, NYISO, MISO and PJM) impose a $1,000/MWh offer cap as part of their overall mitigation measures.

99 The approaches to calculating reference level offers in the various markets is described in https://www.ferc.gov/legal/staff-reports/2014/AD14-14-mitigation-rto-iso-markets.pdf.

Energy market offer mitigation is subject to FERC oversight and carried out by the various internal and external market monitors.

4.8. Incorporation of Renewable Energy Resources

The variable and uncertain output of renewable energy-generating resources can make them a particular challenge when it comes to their incorporation in capacity markets. As these resources become more common, it is important that they be accounted for in the capacity market context. However, special considerations are needed to determine to what extent renewables can be relied upon to support resource adequacy. This section provides an introduction to how renewable resources may be accounted for in capacity markets. This is followed by a discussion of the interaction between capacity markets and the types of public policies (e.g., subsidies and renewable portfolio standards) that have played such a vital role in influencing the generation resource mix towards increased renewable resource penetration. Note that capacity markets are not intended to directly facilitate the achievement of policy targets that, for example, require the development of certain quantities of renewable generation. Rather, capacity markets may play a role in this context by supporting resource adequacy *in spite of* public policy requirements.

4.8.1. Renewable Resource Qualification

It is common practice to allow variable generation, particularly renewables like wind and solar, to participate in the capacity market.[100] This practice recognizes that solar and wind generation have the capability to contribute to reducing system LOLP and, therefore, serve the reliability objective of a capacity market by reducing the likelihood of capacity shortfalls. Renewables' contribution is typically only a relatively small percentage of their installed capacity. The actual percentage of a renewable generator's nameplate capacity that may qualify as capacity will usually depend on season, on production patterns at each generator location, and on the geographical diversity of the resource. For each market, protocols are established to determine how much capacity a resource can qualify and whether any locational or seasonal variation is allowed. Qualified capacity will generally be updated on a regular basis depending on historical performance outcomes. New resources may have their capacity contribution established based on the performance of existing resources of similar vintage and/or existing resources in a similar location.

4.8.2. Interaction Between Market Rules and Public Policies

If variable generation resources are allowed to participate in the capacity market, careful consideration of the interaction between the renewable energy program contracts and the capacity market would be needed to avoid distortion of the capacity market by subsidized resources. Likewise, any other resources that are supported by public policies may have their economics or behavior affected in ways that influence market

100 The exception to this is Great Britain, where renewables are generally excluded because they receive government subsidies. This issue is addressed more fully in the following section.

outcomes. For example, jurisdictions may seek to subsidize local generation resources to maintain associated jobs and tax base, to exercise buyer-side market power and reduce energy costs for local consumers, or to support low-carbon generators that may be important for current or future environmental objectives. In any case, public policies are likely to shift resource economics. In turn, resources may have the incentive (or obligation, possibly as a condition of receiving a subsidy) to offer below cost, thus driving down energy market-clearing prices, potentially below competitive levels. This raises questions about both fairness and the ability of the market to continue to ensure efficient procurement of the level of resources necessary to maintain resource adequacy.

How to address the interplay between policy and capacity markets is a timely subject that is a matter of intense, ongoing debate. Some jurisdictions (e.g., PJM) are considering an approach focused generically on allowing local jurisdictions to fulfill individual policy interests, while also ensuring the fidelity of the price and quantity procured in the capacity market. Other jurisdictions (e.g., ISO-NE) are attempting to look more holistically at how capacity markets can be modified or augmented to achieve both the resource adequacy mission and other public policy objectives. Given that carbon reduction is the most common objective of the relevant public policies, some proposals for reconciling with capacity markets include:

- **Carbon adder to the energy market**: This adder would include the explicit addition to offer prices based on a determined cost of carbon. This would then inform capacity market parameters and participant behavior. In Alberta, the carbon price levy on natural gas should have a similar intended effect.
- **Forward clean energy market:** This supplemental market would provide a centralized, long-term energy auction to allow entities with compliance obligations to procure forward the kind of resources required by policies of each jurisdiction. Such a forward market could operate alongside a capacity market. This is similar in nature to Alberta's Renewable Electricity Program.
- **Bifurcated auction:** Split the capacity market into two stages, one with and one without subsidized resources. This could potentially run alongside other structures, like a forward clean energy market, and would provide two clearing prices, one for merchant resource and one for resources participating on a subsidized basis.

How any of these constructs would operate, and how they would affect and coordinate with capacity markets, is still a matter of debate. Thus, the formulation of a complete set of policy-responsive market rules is still prospective, and any articulation of best practices is likely years away. In the meantime, many jurisdictions have taken an *ad hoc* approach to addressing the effects of public policies, with regular revisions to rules associated with resource qualification and market power mitigation.

4.9. Incorporation of Demand Response

The lack of an active demand side — one that is subject to energy spot prices and is able to respond accordingly — is one of the underlying factors that supports the need for capacity markets.[101] However, improvements in telecommunications capabilities, as well as enabling market design rules, have created circumstances in which some consumers may play a more active role in balancing supply and demand. To the extent that such resources may contribute to reducing peak load and the incumbent reduction in installed capacity requirement, permitting DR to act as a supply resource may increase market efficiency and lower capacity costs. In addition to increasing competition and improving reliability generally, DR resources have the added benefits of being quick to develop, and in certain instances, available at low prices.

Effective incorporation of DR resources into capacity constructs is closely tied to effective development of market rules described elsewhere in this section, including locational pricing, product definition, and above all, performance incentives.[102] From a locational pricing standpoint, the same general principles apply as with traditional supply resources: accurate price signals will promote locational-efficient resource procurement. What constitutes an accurate price signal for DR in electricity markets requires careful consideration. In general, the efficient signal to a consumer to reduce its requirements — whether for energy or capacity — is simply the benefit it receives from not having to pay for that increment of that product. However, due to imperfections in the electricity markets — particularly related to having a lack of transparency to, and participation from, the demand side — consumers and their representatives (i.e., LSEs) are often poorly positioned to express their willingness or desire to curtail load. DR providers are better situated to aggregate and bid load reductions into the market on behalf of customers, and rules in many markets permit this. These bids to curtail are generally submitted on the supply side, as that is a more workable formulation within market constructs that are not well suited to represent demand elasticity. This approach requires that precautions be taken to ensure (1) that DR resources are compensated at a level that is efficient, reflecting the service being provided in a non-distortionary fashion and (2) that market rules ensure that curtailments are not “double counted” (i.e., benefitting once from a reduction in payments for demand and again for compensation related to providing the reduction). In capacity markets, these concerns are often guarded against by paying DR resources on the supply side, but not reducing the

101 See the discussion in Section 2.2.2.

102 As in energy markets, DR participation in capacity markets should be associated with rules to measure and verify the delivery of product when called upon. Often, the rules to validate the delivery of DR product are the same for energy and capacity.

capacity purchase obligations on the demand side of the LSEs, within which the customers providing the curtailment are located.[103]

Discussions of product definition and DR often relate to the physicality of the capacity product. This is particularly relevant in capacity constructs with longer forward periods. In some instances, DR resources have been accused of placing speculative bids into the forward capacity auction, and then buying out of their position in the incremental auctions if prices have not been sufficiently high to subscribe customers. Concern has been registered that such bids are not compatible with the intent of a capacity market, which is a tool to maintain reliability, and therefore, not a forum in which financial speculation should be allowed. To address these issues, a market operator may require more detailed information about the specifics of a DR offer to show that it is a dependable resource, or the operator may hold a DR provider to a more stringent standard should it seek to buy out of its position prior to delivery.

Establishing fair and non-discriminatory performance expectations for DR in capacity markets is also critical. This exercise involves recognizing both the need to procure dependable (and comparable) capacity resources, while also recognizing that DR resources often have inherent limitations. On the one hand, capacity markets that clear a defined product at a single clearing price should ensure that all resources that receive market revenues are capable of providing the same, or at least a very similar, service. On the other hand, while DR is likely part of a cost-effective capacity portfolio, DR resources are often incapable of performing exactly the same tasks as a generator. For example, assets that provide curtailment services may be limited by time of day, season, duration of interruption, and frequency of interruption. Bearing in mind these considerations, market rules should strive to ensure that DR can qualify to participate in capacity markets, while also being sufficiently strict to guard against concerns that DR resources are being provided an equivalent payment to provide an inferior service. System operators or regulators may also determine that it is advisable, particularly in early years, to avoid over-dependence on unproven DR resources to ensure reliability, and may therefore, implement a maximum procurement constraint in the capacity market.

103 Similar considerations apply to DR participating in energy markets. However, unlike in capacity markets, energy market participants who curtail necessarily reduce their metered consumption and therefore avoid the cost of energy associated with that consumption at that time. To pay such resources the full market price to curtail would be an inefficient double payment (that would necessarily result in revenue inadequacy charges in the market and the need to apply uplift charges to maintain a balance of payments). Instead, an efficient payment is the market price less the energy (and possibly transmission) component of the price at the time of curtailment. Providing efficient remuneration for DR in the energy market is important to ensuring an overall compensation scheme across markets (spot and capacity) that leads to efficient levels of investment. For further discussion of this issue, see William W. Hogan, "Demand Response Pricing in Organized Wholesale Markets," May 13, 2010, available at https://www.hks.harvard.edu/fs/whogan/Hogan_IRC_DR_051310.pdf and William W. Hogan, "Providing Incentives for Efficient Demand Response," October 29, 2009, available at https://www.hks.harvard.edu/fs/whogan/Hogan_Demand_Response_102909.pdf.

4.10. Relationship to the Energy and Ancillary Services Markets

As described in detail in Section 3.1.4, capacity markets are inherently administrative constructs designed to address market failures in the energy and ancillary service markets, as well as out-of-market interventions in resource adequacy standards and wholesale market rules, that lead to missing money. Capacity markets seek to competitively determine, and then provide the amount of missing money that must be paid to procure and maintain supply capability to the required level of resource adequacy. The smaller the difference between that revenue requirement and the expected revenue from energy and ancillary service markets, the smaller the sums that need to be paid out by the capacity market for the "right" amount of supply to be available.

For a secondary, administrative market, reducing the amount of payments provided through the construct is a laudable goal. First, from an economic perspective, it is desirable to minimize the need for supplemental revenue beyond that provided in the primary commodity market. Second, from a political and regulatory perspective, capacity markets have proven controversial and difficult to administer. The smaller the payments available, the less time and energy will be spent debating the effects of arcane market rules, and the less effort will be spent by interested parties on (probably) rent seeking through regulatory capture. Therefore, while a capacity market may be found to be necessary to achieve competitive, reliability and/or policy goals, continued effort should be applied to ensure that energy and ancillary service markets are operating as efficiently as possible with as few interventions as feasible. Initiatives may take numerous forms, including improved scarcity pricing, additional demand-side participation, and more accurate price formation in energy and ancillary services. These efforts should, in turn, increase energy and ancillary service revenues, decrease missing money, and reduce the overall scale of capacity markets as an administrative side market.

Moreover, some have argued that capacity markets, once instituted, may eventually be phased out. In some cases, electricity systems have particular characteristics — like a high proportion of hydroelectric power or institutional legacy of conservative reserve margins — that suggest a persistent need for a capacity market. However, in other systems with primarily thermal generation, capacity markets may "be seen as a transitory and optional tool which could ultimately be phased out in the future if markets design improves."[104] That is, for many systems, advances in technology and improvement in market rules may usher in circumstances (e.g., personalized reliability, participating demand side) in which a once-justified

104 See Fabian A. Roques, "Market Design for Generation Adequacy: Healing Causes Rather Than Symptoms," *Utilities Policy*, Vol. 16, No. 3, May 2008.

capacity market may no longer be necessary.[105] Where possible, advancing movement towards such a first-best outcome should be the goal.

105 While it is possible to articulate such a future in theory, it is by no means a certainty. Moreover, no jurisdiction that has implemented a centralized capacity market has ever transitioned away from that capacity mechanism to another, let alone to an energy-only market.