AN ENERGY-ONLY MARKET FOR RESOURCE ADEQUACY IN THE MIDWEST ISO REGION

November 23, 2005

An Energy-Only Market for Resource Adequacy
In the Midwest ISO Region
Table of Contents

AN ENERGY-ONLY MARKET FOR RESOURCE ADEQUACY IN THE MIDWEST ISO REGION

November 23, 2005

INTRODUCTION

This paper continues the discussion of resource adequacy issues initiated in the Midwest through working groups organized by the Midwest Independent System Operator ("ISO") and the Organization of MISO States (OMS).[1] It follows a White Paper prepared by the Midwest ISO and issued on August 3 and September 9 (revised) of this year and comments submitted by stakeholders.[2] In its earlier White Paper, the Midwest ISO expressed a desire to examine the feasibility and merits of using what is commonly called an "energy-only market" approach to ensure resource adequacy. Comments by stakeholders, including important questions asked by the OMS,[3] seek additional information about how an energy-only market would function. In this paper, we attempt to address those questions that can be answered at this time, while providing a framework for thinking about other questions that should be answered as this discussion moves forward.

The principal reason for considering an energy-only market approach to achieving resource adequacy is the expectation that it would allow market incentives, rather than centralized administrative direction, to drive investment decisions. The rationale is consistent with the principal reasons for developing markets and departing from the traditional regulatory structure in the first place. In the words of William Hogan:

> A main feature of the [energy-only] market would be prices determined without either administrative price caps or other interventions that would depress prices below high opportunity costs and leave money missing. The real-time prices of electric energy, and participant actions, including contracting and other hedging strategies in anticipation of these prices, would be the primary drivers of decisions in the market. The principal investment decisions would be made by market participants, and this decentralized process would improve innovation and efficiency. A goal would be to avoid repeating the problem of leaving customers with stranded costs arising from decisions in which the customers had no choice. This change in the investment decision process and the associated reallocation of risk would arguably be the most important benefit that could justify greater reliance on markets and the costs of electricity restructuring. If this were not true, and if it would be easy for planners and regulators to lay out the trajectory of

---

[1] This paper was prepared by and at the direction of Midwest ISO staff, with significant assistance from John Chandley and Robert Borlick. Helpful comments were provided by William Hogan and Mike Cadwalader.

[2] Midwest ISO, "Discussion Paper on Resource Adequacy for the Midwest ISO Energy Markets," posted August 3, 2005; revised, September 9, 2005. http://www.midwestmarket.org/publish/Document/25228f_10631e11216_-7fac0a48324a. Hereafter referred to as "White Paper."

[3] http://www.midwestmarket.org/publish/Document/2b8a32_103ef711180_-77cd0a48324a.

needed investment for the best portfolio of generation, transmission and demand alternatives, then electricity restructuring would not be needed.[4]

From this perspective, the goals of an energy-only market are to improve innovation and efficiency, avoid the problems of stranded costs, and shift the risks and rewards of prudent investment from consumers to investors.[5] This would be achieved by moving the primary responsibility for investment decision-making from regulated planning and centralized administrative processes to the decentralized, voluntary decisions made by market participants responding to prices set by the market. But these are not the only reasons for considering an energy-only market.

One need not be convinced of the innate superiority of competitive markets to recognize other, very practical considerations that favor an energy-only approach to delivering an essential service to consumers without compromising reliability. The structure of an energy-only market directly defines the incentives that induce resources and loads to take actions consistent with reliable operations in the short run. Hence a further justification for giving serious consideration to an energy-only market approach is that it offers a consistent set of incentives that directly support both real-time reliability and resource adequacy. In contrast, current installed capacity (ICAP) structures and reform efforts make clear that achieving equally consistent and effective incentives in a capacity construct is extremely difficult and getting agreement on these administrative mechanisms is even harder, generally resulting in less than adequate incentives in both the short-run and long-run. These are not trivial concerns; they go to the heart of a very difficult problem we are trying to solve.

This paper expands the energy-only discussion by providing a more detailed description of how an energy-only market would work and what the ISO and others would need to do to implement such a market. The paper describes and discusses each element of that market and how it serves in promoting resource adequacy as well as ensuring operational reliability.

Defining An Energy-Only Market

As broadly defined in the earlier White Paper, an energy-only market explicitly pays resources only for the energy and ancillary services they deliver. It does not pay for installed capacity (ICAP). There is no requirement for utilities or load-serving entities to acquire or contract for "capacity" *per se*, or would anyone need to administer markets for

---

4 William W. Hogan, *On an "Energy-Only" Electricity Market Design for Resource Adequacy*, paper prepared for the California ISO, September 23, 2005. See http://ksghome.harvard.edu/~whogan/.

5 The term, energy-only market, is a misnomer and actually refers to a series of closely linked sub-markets for spot energy, operating reserve, other related ancillary services and bilateral contracts. The common bond is that all depend on cost-reflective, transparent spot energy prices in order to function efficiently and effectively.

such "capacity" because they wouldn't exist.[6] This does not mean that energy-only markets do not compensate generators for their "capacity" costs (i.e., their fixed operating costs, start-up and no-load costs, plus the recovery of, and return on invested capital). However, generators recover these costs through the enhanced profit margins (scarcity rents) they earn from selling energy and ancillary services, rather than through direct payments earmarked to recover those costs.

That said, it is worth emphasizing that an energy-only market deals with more than just energy; rather it is a convenient label for a set of market rules that govern the ISO's day-ahead scheduling and real-time dispatch for energy *and* operating reserve.[7] The mandatory capacity rules of ICAP markets presume that these spot prices will not reflect the true system conditions, thus turn to alternative regulatory requirements to provide adequate investment incentives. By contrast, the energy-only market approach presumes that spot prices can be made to reflect operating conditions and provide the right incentives. The expected stream of hourly spot prices for energy and operating reserve provide a foundation for contracts and investment decisions occurring in a series of interdependent markets (*e.g.*, bilateral contracts, derivative markets, *etc.*) complementary to the ISO run spot markets, that will yield the desired level of resource adequacy.

It may seem strange that the real-time spot market should be the key to talking about long-run resource adequacy. Typically, in discussions of resource adequacy or "capacity" constructs these topics are separated from discussion of the real time spot markets. But they are *not* separate and unrelated and it is a serious mistake to separate them. A coherent discussion of resource adequacy must consider how the ISO conducts real-time dispatch and prices energy.

When these topics become separated, the result will likely fall well short of solving the resource adequacy problem and may also undermine short-run system reliability. Indeed, virtually every problem that the Eastern power markets are having with their current ICAP mechanisms can ultimately be traced to the mistake of ignoring or deemphasizing this critical linkage.[8] The hourly spot prices that signal the need for investments to keep existing plants operational and to build new capacity also induce generators to make their existing units available to the dispatch to ensure safe and reliable operations in real time.

---

[6] The ISO leaves open the question of whether it should administer a voluntary forward market for energy, where market participants that are long or short could adjust their positions.

[7] Of the various ancillary services needed to support power system operation, operating reserve is the service most intimately linked with energy production. To maximize economic efficiency the markets for these two products must be closely coordinated.

[8] Scott M. Harvey, "ICAP Systems in the Northeast: Trends and Lessons," September 19, 2005, available at www.caiso.com. The New England LICAP filing (FERC Docket No. ER03-563-030) and the PJM RPM filing (FERC Docket Number EL05-148 *et al.*) both include substantial and remarkably candid critiques of the problems each ISO has experienced with current ICAP approaches, thus necessitating their reform efforts.

Getting the spot price signals right, and making sure that those who make operational and investment decisions are exposed to those prices, is essential. Even if the ISO ultimately adopts a capacity-based construct, that mechanism is unlikely to work very well if the designers lose sight of this basic truth.

The central question raised by the ISO in this paper, and in its original White Paper, is whether an energy-only market would be better than one that includes an explicit capacity mechanism. Would it be easier to implement? Could it do a better job of stimulating investment and ensuring resource adequacy? Would it be easier to design? This paper discusses why the central question is so important and explains why consideration of an energy-only market framework is both necessary and worthwhile no matter which path the Midwest region pursues.

Energy-Only Market And The Need For Intermediate Or Backstop Mechanisms

The energy only market has been characterized by some as an "end state" – a position that the market should eventually evolve into, but one that is not immediately feasible. This raises two issues that must be addressed. First, if an energy only market is the desired end state, then the short and long term impediments need to be identified and either overcome or mitigated. Second, it must be recognized that there is a fundamental difference between a "backstop" mechanism, *i.e.* one that is hopefully never used and is triggered only when a set of pre-defined criteria are met, and an "interim" mechanism which is a step along the way to an energy-only market. In either case it will be necessary to design structures that are incentive compatible with the development of an energy only market. That is, both the "backstop" or "interim" mechanisms cannot provide disincentives for either reaching the end state or ensuring the potential failure of the end state once implemented.

This paper explores the critical issues that need to be addressed to determine the need for an interim capacity mechanism. The paper also summarizes and provides references to recent papers that describe alternative capacity mechanisms.

The Role Of State Regulators

A purely market-based energy-only wholesale market can work without changes in the way electricity is consumed and paid for by end-users but the task becomes much more difficult to achieve. In particular, markets – for any service or commodity – function best when the costs and benefits of specific actions are either implicitly or explicitly transparent. When actions are linked to costs and benefits then parties have the incentive to manage risk in ways that allow for socially optimal outcomes. Under current rules ensuring that adequate capacity is installed is the responsibility of the States. Therefore, to assure both short-run reliability and long-term resource adequacy, wholesale market rules and retail regulation need to work in mutually supportive ways. As a result, state regulators play a critical role in a well functioning energy-only market.

In today's MISO energy markets, market participants are credited or charged based on the locational marginal price (LMP) at their injection or offtake node. Market participants that serve load (LSEs) with sufficient contract cover can insulate themselves from the potential for real-time LMP price volatility. LSEs without sufficient cover through contracts with generation may be able to make up the shortfall through contracts with their customers having price responsive demand. Without either, these LSEs can be exposed to significant real time price volatility. As discussed later, state regulators will have access to LMP data for LSEs under their jurisdiction and will have knowledge of their contract cover, including demand-side contracts. The state regulators can use this information to guide LSE procurement strategies.

Over time retail rate designs can be changed and real-time, interval metering and billing can be implemented. However, decisions need to be made regarding how fast this should be done and for which customers. Because these decisions reside with the state utility commissions an energy-only market cannot be designed and implemented exclusively by the ISO. Important elements, particularly with respect to retail rate design, demand-side response and forward contracting would benefit from significant attention from state regulators. [9]

Regardless of the approach taken to resource adequacy, the energy charge component of any retail rate design should reflect, as closely as practicable, the real-time spot price where the load is located. Any retail rate designs that reflect this locational energy charge, coupled with appropriate metering and billing systems, would charge customers on the basis of cost causality and would facilitate economically efficient demand-side response, which in turn would help define and mitigate prices, discourage market power, and significantly reduce the need for involuntary curtailments (rolling or tailored blackouts). We expand on these concepts in later chapters.

An important issue is whether contracting should be promoted or mandated by state regulators to complement an EOM. In addition to self-supply by utility-owned generation, power contracts of varying terms could play a dominant role in the prices retail customers would face. Such contracts could protect most consumers from the spot price volatility that is essential for an energy-only market to assure operational reliability and long-run resource adequacy. This contracting would not diminish the importance of spot prices in providing the right incentives but it would redistribute the risks of volatile spot prices.

In states where utilities are regulated and have an obligation to serve their franchised customers, such price hedges are a natural consequence of utility plant ownership and power supply contracting. In states that allow retail choice, or where utilities have divested their generation, state regulators would want to consider how best to structure "default service."

The degree of support the ISO receives from state regulators, as well as the degree of cooperation among them (such as by acting collectively through the OMS) would largely determine how successful an energy-only market would be. It is unlikely that the

[9] These structural flaws also interfere with the efficient functioning of capacity-based resource adequacy mechanisms so correcting them is important regardless of what the ISO decides to do.

ISO can establish and implement a fully functional energy-only market design without substantial state cooperation, although the market would still function better if the ISO implemented the key elements of that market in any event. Of course, the same need for support and cooperation would apply for any "capacity-based" mechanism that the ISO might pursue.

Compatibility With PJM's Reliability Pricing Mechanism

The Midwest ISO market shares interfaces with several other markets, including PJM. Six states in the Midwest have utilities and transmission systems that overlap both the Midwest ISO and PJM footprints. In addition, all of the utility systems in either RTO would be affected through their interconnection with the other RTO. Thus, all stakeholders within the two footprints have a vested interest in "seamless" trading between the two RTOs, and most support the goal of achieving a "Joint and Common Market."

Several questions arise from the Midwest ISO-PJM interconnection. First, is it necessary for the two RTOs to employ similar resource adequacy mechanisms to assure seamless trading? Second, if their market designs proceed along different paths how can they be made compatible? Third, would inter-RTO trading suffer if they were different, and if so, how could those problems be mitigated?

PJM has recently proposed to address resource adequacy issues in its footprint through what it calls a "Reliability Pricing Mechanism" (RPM), which is now before the Federal Energy Regulatory Commission. RPM is a very different mechanism for achieving resource adequacy than an EOM; it is a capacity-based construct employing an innovative "capacity demand curve." This novel feature has diverted attention away from the important linkage between real-time spot prices and resource adequacy.

Given the tentative status of the RPM filing, this paper cannot fully answer all of the questions raised earlier; however, it does provide an analytical foundation for assessing the issue.

I. THE ENERGY-ONLY MARKET

The defining characteristic of an energy-only market is that prices are set by supply and demand while minimizing the use of administrative caps or other means that artificially suppress those prices below competitive levels. Although some administrative intervention may be needed to deal with missing markets, such as those for some ancillary services or other possible forms of "market failure," one would choose a solution that would least distort spot energy prices.

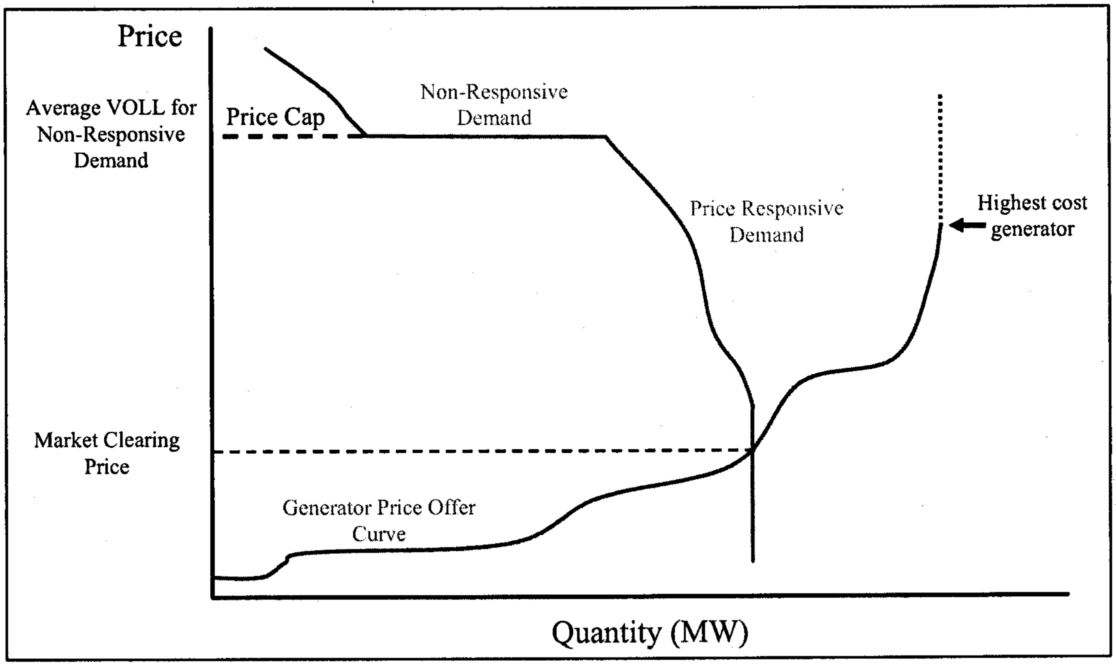
HOW ENERGY PRICES ARE DETERMINED

The common notion that an energy-only market equates to *unbridled pricing and "let-the-market-rip" is **not** what is meant here by an energy-only market.* Nor does it mean that prices can reach infinite or indeterminate levels during supply shortage conditions. In an energy-only market, there are constraints on prices.

One constraint is competition among generators, i.e., market power is minimal or mitigated (this issue is addressed in a later chapter). A second constraint is the ability of consumers with flexible loads to reduce their consumption in response to high spot market prices. Finally there is an overall cap on spot energy prices reflecting the willingness to pay of those consumers who cannot respond in real time to the spot price. These three price constraints are illustrated in the figures that follow.

Figure 1 shows how the spot energy price is determined when available generating capacity is sufficient to satisfy demand.

Figure 1. Price Determination When Supply is Sufficient [10]



The spot energy price is set by the intersection of the generation price offer (supply) curve with the demand curve. Assuming there is sufficient competition among the generators the supply curve will closely track the generators' short-run marginal costs of production, which is almost all fuel expense. Thus, this first constraint is embedded in the generation price offer curve.

The second constraint on energy prices is the price-responsiveness of customer loads, represented by the downward sloping portions of the demand curve. The rightmost

[10] The portion of the demand curve labeled "Non-Responsive Demand" indicates the amount of load that end-use customers have that is either not sensitive to price, or represents load where customers may not see the wholesale price. Usually economists indicate demand not sensitive to price as vertical segments; here it is drawn horizontally to indicate prices are capped at the average VOLL.

segment of this curve represents loads that value electricity at less than the "average VOLL." The leftmost downward sloping segment of the demand curve represents price-responsive loads that value electricity by more than the "average VOLL" price.

The horizontal portion of the demand curve represents the loads of customers that are not price-responsive, primarily because they lack real time meters thus are not billed on an hourly basis. Although their loads value electricity across a broad spectrum of electricity prices, they are all represented here at the average valuation for the entire load group because the customers cannot directly reveal the valuations of their individual loads by turning them off or on in response to spot energy prices. Thus, "average VOLL" price is defined to be average value of lost load (VOLL) for this group of spot price-insensitive customers when electric service to them is randomly interrupted.

This brings us to the third constraint on spot energy prices. As illustrated in Figure 1, the market design imposes an overall cap on real time spot energy prices equal to the average VOLL. This absolute price ceiling prevents the market from inefficiently purchasing electric energy at prices that would exceed the prices the customers incrementally served by that energy would be willing to pay.[11]

It is likely that the vast majority of customers, including virtually all residential and most small C & I customers, would be part of the horizontal segment of the demand curve. A defining characteristic of these non-responsive consumers is that they lack real time metering, which reflects the realities of how these customers are metered and billed today. Furthermore, in most cases it would not be cost-beneficial to install these meters because the savings they could realize through voluntary load curtailments would not recover the cost of the requisite real time metering and billing.

For this large middle group of customers, it can be assumed that changes in spot energy prices would have essentially no effect on their short-run usage – hence their portion of the demand curve is flat. They would continue to make their energy consumption decisions based on their tariff rates, without regard to the level of spot prices, even if those spot prices occasionally rose to very high levels.[12]

[11] The incrementally served customers are those whose loads would not be curtailed if additional electric energy were available at prices above average VOLL and were purchased to serve them. Such high cost energy would likely come from generators that have marginal costs well below average VOLL price but that seldom run, thus can only recover their fixed costs by charging very high prices during the few hours they do run. These generators need not be technically inefficient (i.e. high heat rates, low availability, etc); indeed, they could be state of the art plant. However, if there are too many of them, relative to customer demand, each will only run a small number of hours per year, which means they are economically inefficient and everyone would be better off if they had not been built. One function of the price cap it to ensure that consumers do not subsidize the construction of excess capacity.

[12] However, if spot energy prices remained high for extended periods, as they did in the 2000-01 California electricity crisis, consumers billed under fixed tariffs would still see high monthly bills which would in the longer-run would induce them to inefficiently reduce their usage in both high priced and low priced hours.

As stated earlier, we don't know how much the price-insensitive customers would be willing to pay to avoid involuntary curtailments. Nonetheless we must assume some value for their average VOLL in order to set the price cap. In resource adequacy discussions in the US, the estimated averages range from about $2,000 to $10,000 per MWh or more, depending on which categories of customers are assumed to be curtailed.

The prices paid by price-responsive customers would never exceed their willingness to pay, as determined by their own choices to consume or not. In principle, the prices paid by non-responsive demand would also never exceed their willingness to pay if the market price cap accurately reflects their average VOLL.

Customers with price-responsive loads would be served under retail rate designs employing energy charges that closely track the real-time hourly wholesale spot energy prices.[13] In addition, their usage would be metered on at least an hourly interval basis and their responses to the hourly spot prices would be properly reflected in their bills. The price-responsive loads need not be dispatchable by the ISO or by the customers' load serving entities. All that is essential is that these customers be able to respond to the spot prices by voluntarily curtailing their consumption or shifting it to lower-priced hours.

There is no strict, minimum requirement for the amount of price-responsive load needed to achieve optimal resource adequacy. This is an issue that must be empirically assessed, as discussed in a later chapter. Obviously, the more price-responsive load there is the more economically efficient the market will be and the lower the total costs.[14] However, a subset of the largest industrial customers facing real-time energy prices is likely to be all that is required to control market power and reduce costs for all customers. During conditions of scarcity, even a modest amount of price-responsive demand could dramatically limit the level of electricity prices for all customers while also forestalling rotating blackouts.

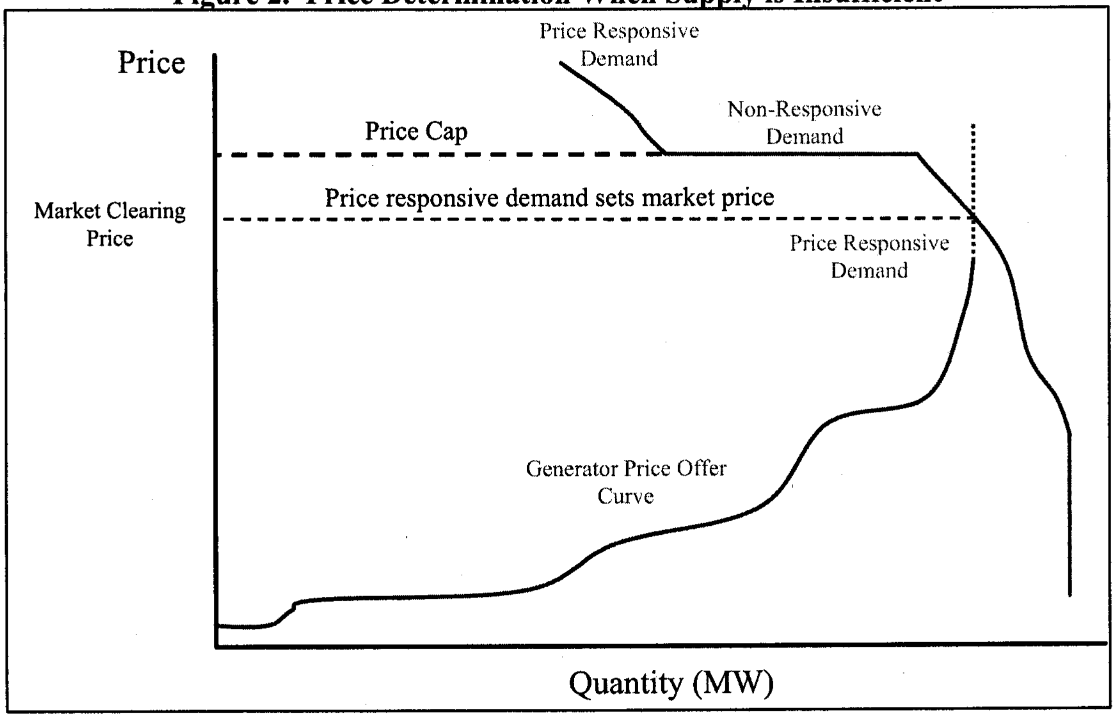
The earlier figure illustrated how the energy price is determined when available

[13] In fact it is quite likely that most price responsive customers have loads in both value categories.

[14] If more demand is price responsive, less generation capacity will be needed to meet the desired loss-of-load expectation adopted for the non-price responsive loads. Conversely, if few consumers are price responsive, the total costs of reliably meeting all loads will be higher. In deciding how much price-responsive demand to encourage, state regulators, regardless of their views on "restructuring," should keep this fundamental tradeoff in mind.

generating capacity is abundant. Figure 2 below illustrates what happens when generating capacity is in short supply and price responsive demand fills the shortfall.

Figure 2. Price Determination When Supply is Insufficient



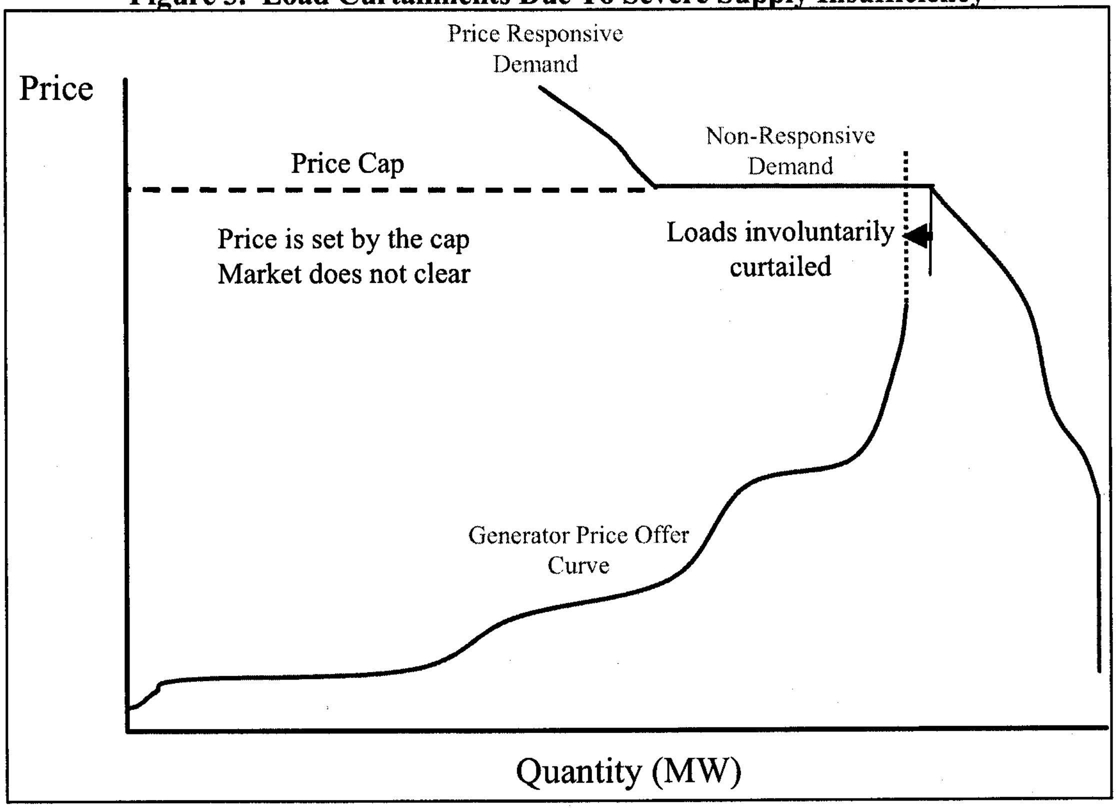
In this situation the spot energy price is determined by the demand curve intersecting the generation supply curve at maximum available capacity, which can occur at a market price substantially higher than the short-run marginal cost of any generator. This pricing rule is referred to as "scarcity pricing" or "shortage cost pricing" and it is a critically important feature of the energy-only market design.

Figure 3 illustrates what happens when supply insufficiency is so severe that involuntary curtailments result. As shown there, the spot energy price rises to the cap and some non-price responsive customers must be curtailed, perhaps through rolling blackouts or based on which customers had inadequate contract cover prior to the operating day. From a societal perspective the first customers to be interrupted should be those that place the lowest monetary value on electricity. Value of service studies consistently suggest that residential customers value electricity far less than commercial and industrial customers do.[15] While it may appear inequitable for residential customers to always be interrupted first, we must remember that the customers remaining on the system pay very high prices for their consumption during hours of stress and substantially contribute to the fixed costs of the generating capacity serving them. Their payments

[15] Lawton, Sullivan, VanLiere, Katz and Eto, A Framework and Review of Customer Outages: Integration and Analysis of Electric Utility Outage Cost Surveys, Report LBLN-54365, Lawrence Berkeley National Laboratory, November 2003.

make it possible for that generating capacity to be economically viable and therefore available to serve the curtailed customers at other times.

Figure 3. Load Curtailments Due To Severe Supply Insufficiency



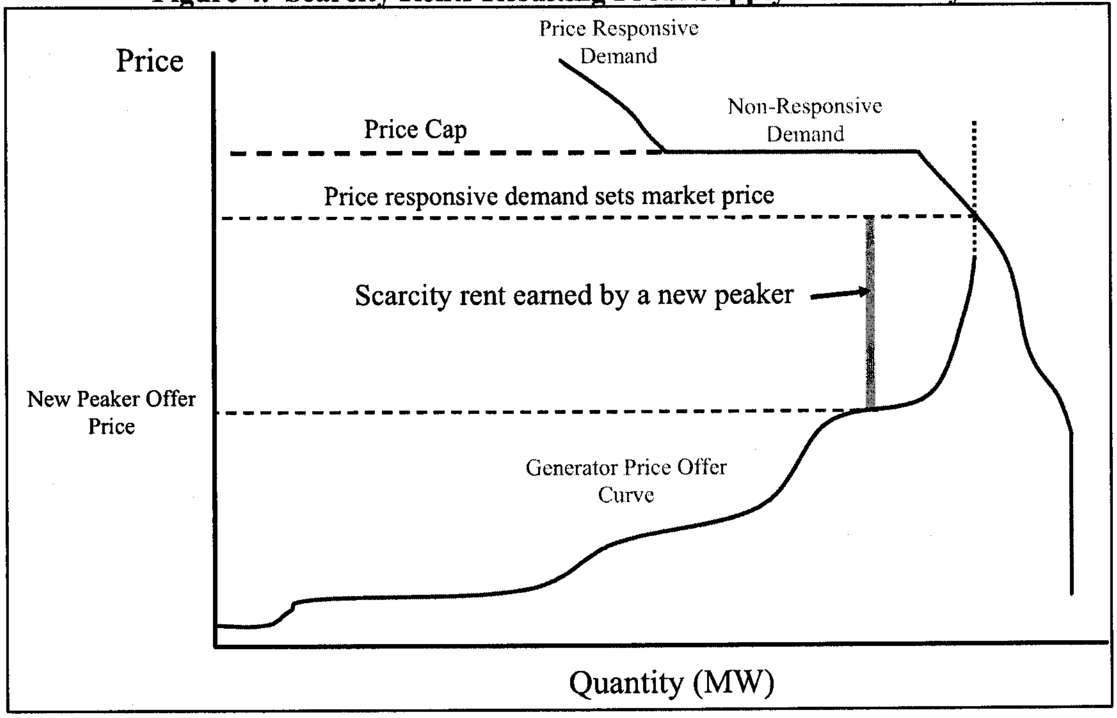
The Role of Scarcity Rents

Let us return to the question of how scarcity pricing in an energy-only market can assure resource adequacy without the need for capacity payments.

Figure 4 illustrates how scarcity pricing produces "scarcity rents" for all generators operating in hours when such pricing occurs and these rents contribute to the generators' fixed costs, capital recovery and profits.[16] For clarity of explanation the figure only highlights the rents for one plant – a new peaker – but all of the other plants producing energy also earn scarcity revenues. It is the expectation of earning such scarcity rents that produces the incentive for owners of existing generating plants to make those plants available for dispatch when the market needs them most, and also the incentive for developers to invest in new generating plants.

[16] As used here the term, "scarcity rent" means the difference between a generator's sales revenues and its variable operating costs (which exclude any startup or no-load costs) during a given time period.

Figure 4. Scarcity Rents Resulting From Supply Insufficiency



The latter incentive is depicted by the shaded area depicting the scarcity rent that a new peaking unit would earn. If developers expect a new peaking unit to earn sufficient scarcity rents to cover its fixed costs, recover its capital investment and earn a compensatory rate of return on that capital, revenue adequacy will be assured. Thus, in an energy-only market scarcity rents are the genesis of resource adequacy as well as operational reliability.

The specific process through which scarcity rents work to promote resource adequacy is as follows. Assume that we begin with the power system in a state of overcapacity (as is the case today throughout the MISO footprint). As demand grows, the generation reserve margin shrinks and price-responsive demand will be called on with increasing frequency to compensate for shortfalls in generating capacity. The more frequently price-responsive demand sets the energy price the greater will be the scarcity rents and the more profits generators will earn, *and will expect to earn in the future*. This process will continue until investors' expectations regarding future scarcity rents reaches a level that they accept as sufficient to justify constructing new peaking capacity. As the new plants enter service they will arrest further declines in the generation reserve margin and maintain it at that level.

For purposes of achieving generation resource adequacy all that is required is the economic viability of new peaking units. However, once such viability is assured, investors will also have incentives to build other types of plant, driving the power system toward the optimal mix of generation fuels and technologies. This will happen through the following process.

As load grows and new peaking units are added to meet peak demand, the most fuel-efficient units will be dispatched to run more and more hours each year. Eventually the point will be reached where a different generation technology, such as a combined-cycle unit, can produce energy at a lower average cost when run at these high capacity factors, producing higher profits for its owners. This same argument also extends all the way down to base load unit additions.

The level of resource adequacy associated with the equilibrium condition just described depends on a number of factors, primarily how high the energy price cap is set, how much price-responsive demand exists, how much are the capital and fixed operating costs of a new peaker, and investors' risk aversion.[17] This issue cannot be resolved through qualitative arguments alone; consequently, we intend to quantitatively address it in future documents.

DEMAND FOR AND PRICING OF OPERATING RESERVE

The supply-demand analyses presented in Figures 1 through 4 are useful (albeit simplified) abstractions. In real power systems generators must supply not only energy to customer loads but also reserve capacity (aptly termed "operating reserve") that can quickly deliver energy in response to a "contingency", i.e., a sudden, unexpected failure of a major generator, transmission line, transformer or other key facility. Thus, the available generation must satisfy a combined demand consisting of customer loads plus the required operating reserve.

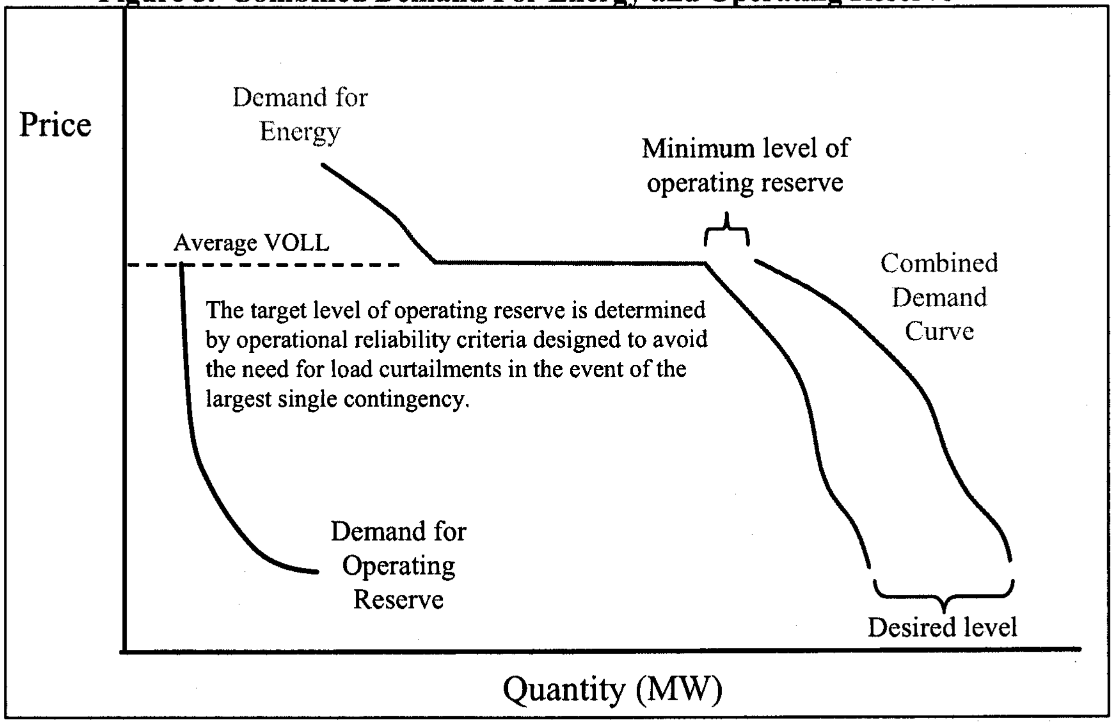
From this it follows that the market price of energy is more closely approximated by the intersection of the generation supply curve and the combined demand curve.[18] Even so, the concepts presented earlier remain valid.

[17] Even with no price-responsive demand an energy-only market would deliver an equilibrium level of resource adequacy but it would likely be unacceptably low by current industry standards. Without price-responsive demand load curtailments would be more frequent, driving the spot energy price to the price cap whenever they occur. The more hours during which this occurs, the more scarcity rents generators earn. Eventually equilibrium would be reached when these scarcity rents are sufficient to attract new generation.

[18] This graphical method is still only approximate but it is useful. Only the effect of "in the money" operating reserve, i.e., that which could profitably produce and sell energy if it were not providing reserve, affects energy prices. Often there are units operating at minimum load in off-peak hours that can provide spinning reserve at no opportunity cost. Also quick-start units, such as pumped storage hydro, can provide non-spinning reserve at no cost. Neither of these effects are captured in simple supply-demand curve diagrams.

Figure 5 illustrates the combined demand curve. As shown there, total demand is displaced to the right of the energy demand curve by the amount of operating reserve required. It also shows that as the energy price increases the system operator will trade operating reserve for energy until some operating reserve minimum is reached, an important concept discussed below.

Figure 5. Combined Demand For Energy and Operating Reserve



Operating reserve comes in two forms:

- spinning reserve and
- non-spinning reserve.

Spinning reserve is that portion of an operating generator's capacity that is not delivering power to the system. This unloaded capacity can begin delivering replacement energy almost instantly in response to a contingency and the amount of lost capacity it can replace is determined by the thermo-mechanical ramp rate of the generator.[19] Spinning reserve is very important in arresting the decay of system frequency and returning it to its target 60 cycle value. A large decline in system frequency would trip

[19] In US power systems operating reserve is defined as the total output that a provider is capable of delivering within 10 minutes following a contingency. For small systems with few or no connections, such as Singapore or New Zealand, the deliverability criteria are measured in seconds, not minutes.

the circuit breakers of operating generators, further compounding the under-frequency problem and ultimately leading to uncontrolled, cascading blackouts.

Non-spinning reserve is provided by generating plants that are off-line but can quickly start up. Hydro units with storage are ideal for this type of service. The only requirement is that they must be able to deliver substantial amounts of energy within the prescribed 10 minute period.

An important feature of an energy-only market is the trade off between energy and operating reserve alluded to earlier when introducing Figure 5. As a general rule, with current technology the demand for operating reserves cannot be defined solely by the market; it is an administrative construct determined by the ISO or by some entity responsible for setting reliability standards, such as the Regional Reliability Council.[20]

Figure 6 shows how the demand for operating reserve is determined as a function of energy price. The curve is concave, reflecting the fact that the marginal value of operating reserve rapidly declines as more and more is available. The ISO should substitute operating reserve for energy as the marginal value of energy to the system increases. However, there is some minimum level of operating reserve which the system operator will not violate, illustrated here as 3 percent of the contemporaneous energy load.[21] Rather than encroach upon this minimum level the system operator will shed load in order to preserve this minimum requirement.

The amount of operating reserve consumed by a contingency is randomly distributed and can be small or large. Obviously the less operating reserve the system has the greater the probability that the system operator will need to shed load. Thus, this probability directly translates into a loss of load expectation (LOLE) for the system. Multiplying this LOLE by the average VOLL applicable to the customers likely to be curtailed, produces the concave expected cost function shown in Figure 6.

Although the system operator would like to have some target level of operating reserve, *e.g.*, 7 percent, it would not maintain that target at any price. As supply tightens, increasing the spot energy price, the price of operating reserve will rise in lockstep and the system operator will trade operating reserve for energy until the minimum operating reserve requirement is reached, at which point the spot energy price will equal average VOLL.

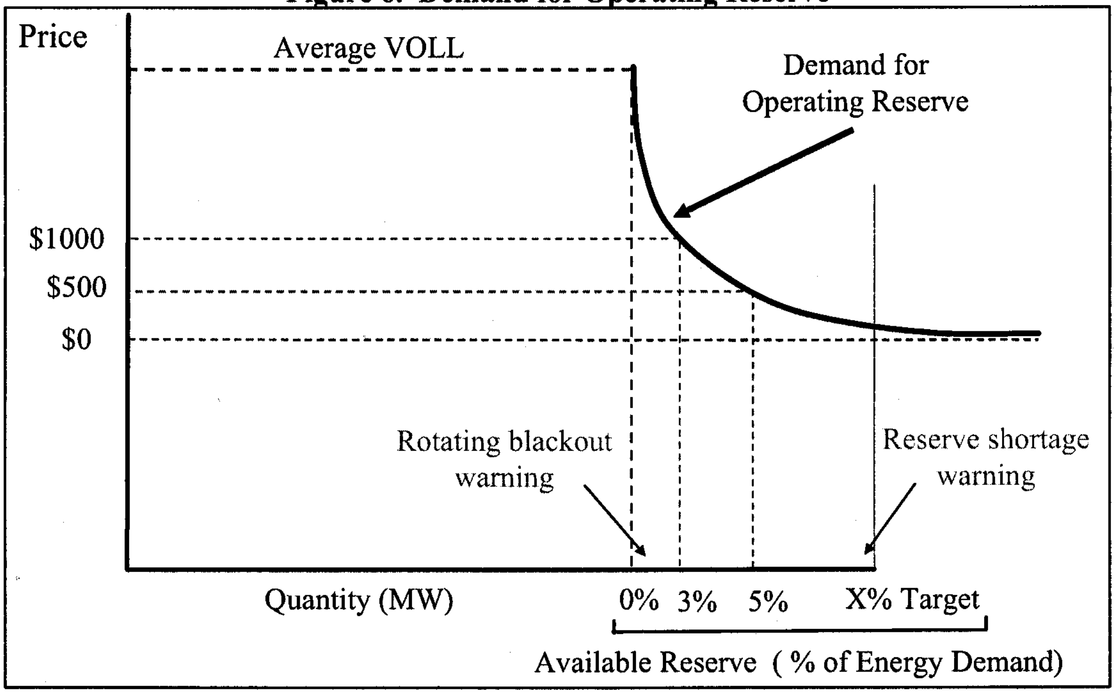
The economic tradeoff between energy and operating reserve produces a combined demand curve that is more steeply sloped than the energy demand curve, thereby producing greater scarcity rents for generators when loads are not curtailed than

[20] In the Midwest ISO, the operating reserve standard varies by RRO; for example, in ECAR it is 3%.

[21] However, for a power system that has few internal transmission constraints and has substantial import capacity, operating reserve can actually be allowed to drop to zero without endangering operational reliability. If done frequently, such free rider behavior would be frowned upon.

would be the case if this operating reserve was held constant. This assists an energy-only market in achieving resource adequacy.

Figure 6. Demand for Operating Reserve



If there's an insignificant amount of price-responsive demand the demand for operating reserve will still produce a downward sloping combined demand curve (albeit a steep one) that produces additional scarcity rents for generators so long as energy prices are determined by the intersection of the supply curve with the combined demand curve. *Again, we emphasize that this approach to pricing spot energy will not occur unless explicit provision for doing so is included in the market rules.*

Optimizing Energy and Operating Reserve Production

The Midwest ISO's economic dispatch of generation currently does not jointly optimize ("co-optimize") the production of energy and operating reserve. Consequently, spot energy prices currently do not reflect the effect of operating reserve shortages if and when they occur. This would change with the adoption of the energy-only market. The economic dispatch process would direct generators to produce energy and/or operating reserve based on their respective energy offer prices, to provide customers with the needed energy or operating reserve at the lowest bid-based cost.[22]

Through such co-optimization, the spot prices paid for energy and operating reserve would be mutually consistent. A generator that is available to provide energy at a given offer price might be directed to partially load its unit in order to provide some operating reserve. The generator would then be paid the spot energy price for its energy and the spot operating reserve price for its operating reserve. The market price for operating reserve would be set by the generator with the largest opportunity cost (i.e., foregone profit from energy sales) thereby ensuring that all other generators would be adequately compensated.

When generating capacity is in ample supply, spot energy prices would be low to modest and prices for operating reserves would be even lower. However, when generating capacity is in short supply, spot energy prices would be very high and the spot prices for operating reserves, though generally still lower, would track energy prices upward. Ultimately, if loads must be curtailed both spot prices would equal the energy price cap.

II. OPERATIONAL RELIABILTY, RESOURCE ADEQUACY AND ENERGY MARKET FAILURE

THE MISO SPOT MARKETS

Since April 1, 2005, the Midwest ISO has operated two spot markets for energy – The Day-Ahead Market and the Real-Time Market. These two markets arise as a natural consequence of its real-time, security-constrained economic dispatch of the generating plants under its control.[23] The Day-Ahead Market sets *ex ante* energy prices for generators and loads bidding into that market. The Real-Time Market sets *ex post* energy prices for generators and loads that do not participate in the Day-Ahead Market or that depart from their commitments scheduled previously in that market. Along with the ISO-coordinated provisions for operating reserve and regulation service by market participants, the ISO's security-constrained economic dispatch ensures safe and reliable system operations every moment of every day across the region.[24]

[22] See the Appendix for a more detailed discussion of the co-optimization of energy and operating reserve markets, with examples of how prices would be determined for energy and operating reserves.

[23] This economic dispatch process adjusts the outputs of all generators under MISO control at five minute intervals in order to supply all customer loads while minimizing the total payments made to generators and also satisfying system security constraints. These constraints prescribe that certain generators and transmission facilities remain partially loaded in order to provide reserve energy in the event of generation or transmission contingencies.

[24] Regulation is the ancillary service that matches generation to load within each hour. It is provided by generators that continuously adjust output in response to small changes in system frequency.

Like the Eastern ISOs, the MISO spot markets employ locational marginal pricing (LMP) to signal the value of energy at each injection and offtake node on the transmission grid. LMPs in the Day-Ahead Market are forecasts of the next-day values produced by iterative computer simulations of the entire MISO system. These prices are inputs to the generation scheduling and unit commitment process. LMPs in the Real-Time Market are *ex post* values reflecting the actual real-time dispatch. Both sets of prices provide accurate signals to generators about the actions they need to take at each location to satisfy customer loads, maintain voltage and manage transmission congestion. In addition, these same prices also lay the foundation for resource adequacy.

The economic dispatch process and its associated spot markets are not just tools for ensuring short-run reliability; they also play an important role in the assurance of long-term resource adequacy. If these spot markets are properly designed the resulting spot prices will reflect the value of the last MW of load served. These price signals will, in turn, encourage generators to make their capacity available when needed and investors to finance the new generation needed in the future. Thus, these two spot markets play a pivotal role in assuring both short-run reliability and long-run resource adequacy.

SPOT PRICES EFFECT GENERATOR AND CUSTOMER BEHAVIOR

For existing plants, the hourly LMP prices for energy (and operating reserve when these markets exist) signal plant owners to take the actions needed to assure short-run operational reliability. The spot prices provide incentives for such actions as performing or postponing maintenance, securing adequate fuel supplies, deciding between spot purchases or firm fuel contracts, procuring fuel storage, maintaining operational crews during key periods and other activities needed to keep the existing plants operable and available for dispatch.

For new plants, the expected future prices signal to investors when and where new capacity is likely to be needed, what types of and sizes to build, what fuels to use, and what capabilities the plant should have (such as quick start, cycling capability, and ramp rates). Although many factors influence investment decisions, expectations of future spot prices are the key driver. Even if investment decisions are premised on long-term contracts, particularly when regulatory and/or market uncertainty is high, those contracts will reflect both parties' expectations regarding spot prices over time.

Spot prices are important regardless of whether the participants decide how to operate their plants or are responding to instructions from the ISO. When such choices are independent, it is critical that the spot prices signal the actions that best support reliability and efficiency. When plants follow instructions from the ISO, consistent spot pricing ensure that the instructions do not act against participants' self interests.

Energy spot prices are just as important on the demand side as they are on the supply side. In the short-run spot prices can signal both consumers and their load-serving entities (LSEs) about whether and when to use more or less energy. In the long run, these same prices influence consumers' decisions to invest in efficiency improvements and other demand-side measures that can change their future energy usage. These actions, in turn, affect not only peak demands but also load shapes.

IMPORTANCE OF SCARCITY PRICING

The concept of scarcity pricing and the scarcity rents it produces was introduced and illustrated in Chapter I. During times when generation is in short supply it is important that spot energy prices be allowed to rise to levels that reflect this scarcity condition. During such times spot prices should be determined by the willingness of consumers to pay for energy, rather than by the energy offer prices of generators. Such relatively high spot prices will yield scarcity rents for generators and cost savings for price-responsive customers while efficiently rationing the available energy to those end uses that value it most highly.

The extraordinary efforts taken by both suppliers and consumers in responding to these high prices will combine to obviate the need for involuntary load curtailments in all but the most extreme capacity-short situations. *The lights will stay on, and no customer will be involuntarily curtailed given their willingness to pay the high spot energy prices.*

Over time, the level of investment induced by scarcity pricing will tend to fluctuate around the level of resource adequacy for which consumers (or their LSEs) have indicated a willingness to pay. But the key feature that allows this mechanism to work is the ability of spot prices to occasionally rise to levels reflecting the actual shortage conditions. *Without such occasional shortage-cost prices the mechanism will fail to produce the desired result.* Some generators will not recover their fixed costs, and over time, investors will support a lower level of investment than that which satisfies the region's adequacy criterion.[25]

MARKET FAILURE FROM SUPPRESSING ENERGY PRICES

Despite the critical importance of scarcity pricing, spot prices are routinely suppressed in every US electricity market. This may occur through caps on generator offer prices or through more general caps on market energy prices. In the Eastern ISO markets and the Midwest ISO, the cap on generator offer prices is currently $1,000/MWh. In California the cap is even lower – currently $250/MWh. It is not the offer caps that are the problem, but rather how they interact with other market rules for determining market prices. Spot prices can exceed the offer caps through other means but the offer caps applied under shortage conditions tend to limit prices below market-

[25] The discussion in this section is only a summary of the economic argument. For more detailed explanations of the economic theory, a useful source is Steve Stoft, *Power System Economics, Designing Markets for Electricity,* Wiley–IEEE Press, 2002.

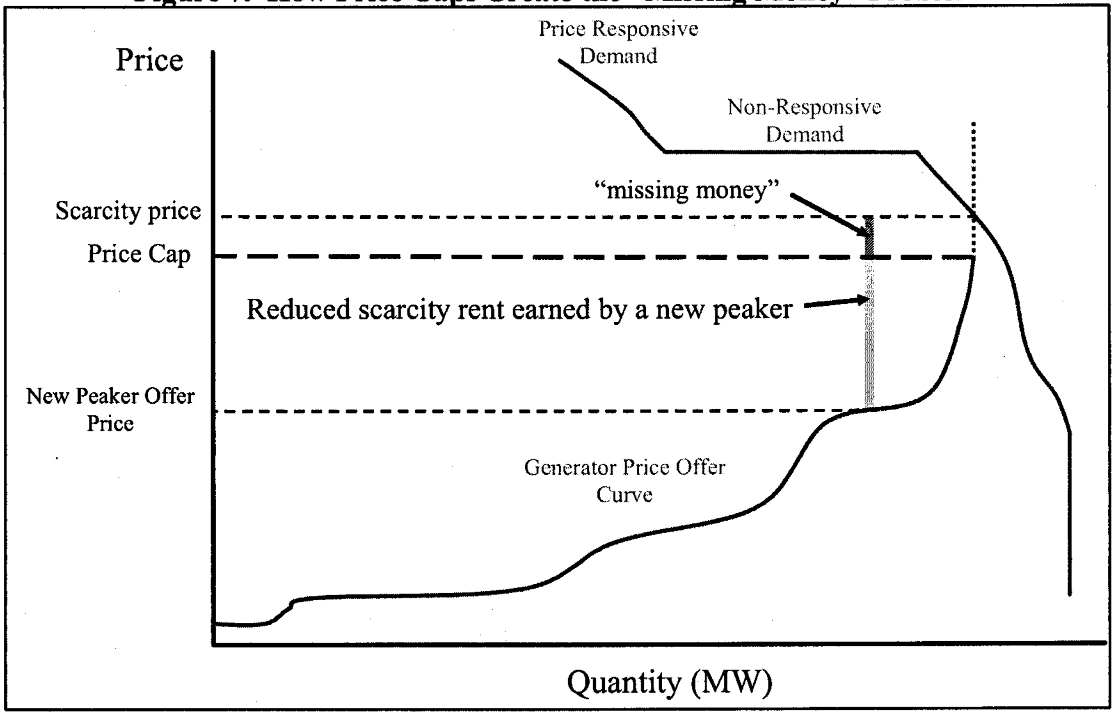
based levels.[26]

Suppressing spot energy prices through price caps and other market flaws, creates the "missing money" problem illustrated in Figure 7. This figure is the same as Figure 3 in Chapter 1 except that the price cap is set lower than the market clearing price.

Shown here by the black diagonal lines is the scarcity rent eliminated by the price cap that a new peaker would otherwise have earned at the market clearing price. This lost scarcity rent is termed the "missing money."

Without the missing money investors will not build or maintain enough capacity to meet the region's resource adequacy goals. Thus, the rationale underlying all capacity payment mechanisms is to replace the missing money.

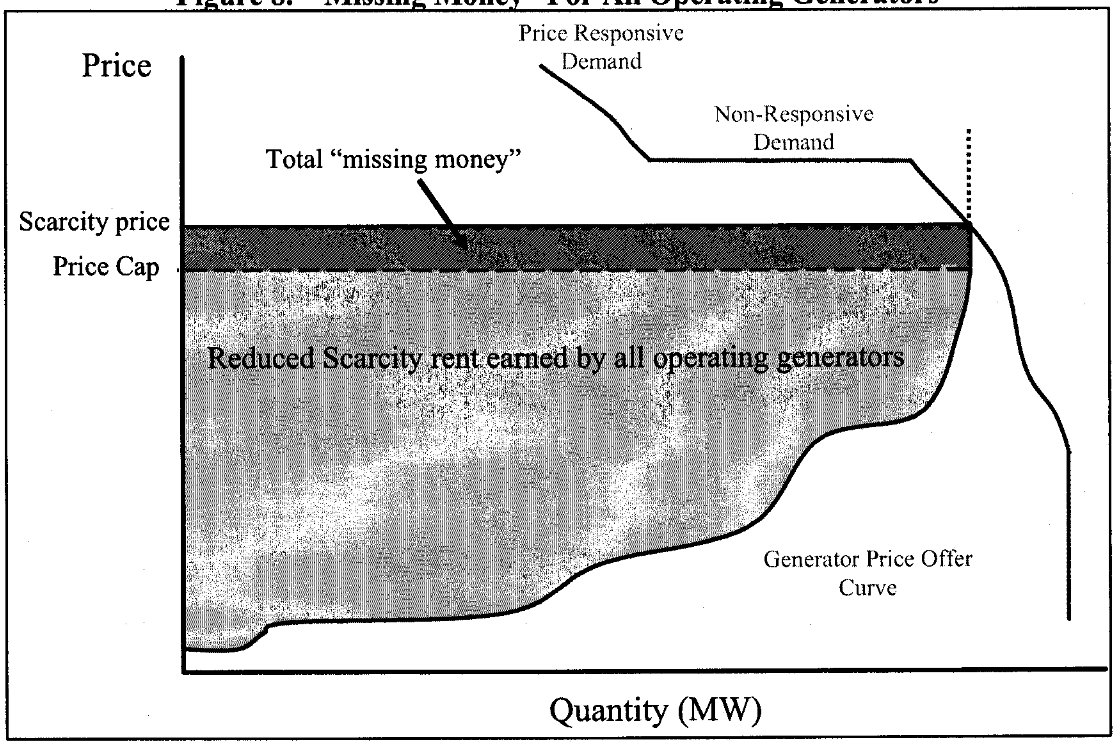
Figure 7. How Price Caps Create the "Missing Money" Problem



[26]The New York and New England ISOs have a modified version of scarcity pricing, which raises the spot prices above the generators' nominal marginal operating costs whenever the ISO experiences shortages in operating reserves; the greater the shortage, the higher the price rise, up to a cap. The shortage-cost adder is determined from an administratively defined demand curve for operating reserves. Such a mechanism would likely be a component in an energy-only market construct, as discussed in the previous chapter.

While figure 7 only illustrates the missing money problem for the one peaker, the price cap reduces the scarcity rents of *all operating* generators as shown in Figure 8.

Figure 8. "Missing Money" For All Operating Generators



Most ISOs, including the Midwest ISO, employ focused mitigation rules to limit generation offer prices when conditions suggest the possibility that market power could be exercised, such as in local areas with significant transmission constraints. Also, generators needed for out-of-merit dispatch to resolve congestion typically have their price offers mitigated. Even when market power is not a concern, spot prices may be suppressed because of political opposition to high spot energy price spikes.

Occasional price spikes are normal and are not *by themselves* indicative of market failure; rather they are natural outcomes in electricity markets that reflect the effects of varying demand, generating plant operating constraints and transmission constraints. Indeed, price spikes are a necessary signal for generators to make extraordinary efforts to be available or to increase output to the maximum. On the other hand, sustained prices at very high levels are not normal; more likely they indicate presence of market power or some market design or regulatory flaw.[27]

[27] During the California energy crisis, several thousand megawatts of existing capacity refused to schedule or offer their power for dispatch because state regulatory decisions prevented the utilities from paying either suppliers under QF contracts or suppliers that might otherwise have sold spot energy through the dispatch. The resulting shortages were regulatory artifacts which contributed to rotating blackouts in early 2001.

Spot prices can also be artificially depressed by ISO dispatch and pricing rules that have little to do with market power or politics. For example, an ISO may commit too many high-cost units to run at minimum levels, or hold high cost units in contingency reserve, while not allowing those plants to affect the LMP spot prices used for settlements.

Another common example is the overuse of "reliability must run" (RMR) contracts, which are often used in locally constrained regions to maintain a plant "needed for local reliability" that might otherwise retire because of depressed spot prices. RMR plants tend to be high cost, but they typically are not allowed to set market prices, which keeps down the market prices in constrained regions that are already short of supply. These, and other administrative mechanisms, suppress spot prices below the levels needed to provide the very incentives to solve the problem.

An energy-only market would require the system operator to carefully review its rules to ensure they do not artificially depress spot prices below competitive levels. If spot prices fail to give the right operational incentives, the system operator will be forced to impose additional administrative rules and sanctions to ensure operational reliability.

Capacity Payment Mechanisms

It is now conventional wisdom that the main purpose of capacity payment mechanisms is to replace the "missing money" that scarcity rents would have provided if spot prices (or contracts based on expectations regarding spot prices) were not artificially suppressed. When an ISO gives up on getting spot market prices right through better pricing of reserves, scarcity pricing, and other market design features, it is forced to solve the missing money problem with capacity payments. In addition, it must turn to non-market payments and administrative rules and sanctions to replace the operational incentives that efficient spot market prices would have provided. This inexorably leads to the need for further administrative rules and sanctions, along with demands for increasingly complex capacity markets administered by the ISO.

Capacity payment mechanisms can take different forms. The most common form is to impose an installed capacity (ICAP) requirement on LSEs (as found in the Northeast and proposed in California). Another approach is to require the LSEs to purchase a prescribed amount of capacity through ISO-coordinated capacity auctions (as in Eastern ISOs). These ISO-administered mechanisms may be short-run (such as monthly, as in New York and New England) or long-run (e.g., for delivery four-years out, as in the PJM RPM proposal). As the acquisition process looks further out, there must be periodic true-ups and interim mechanisms to account for forecast errors or other changes in conditions such as shifting LSE loads and capacity obligations under retail choice.

Once the ISO introduces a forward auction for generation resources, the natural instinct is to allow alternatives, demand-side resources and transmission upgrades, to compete on the same playing field. In fact, without appropriate scarcity pricing in the spot-market, explicit subsidies to demand-side and transmission alternatives would be logical. The tradeoffs between these options are quite complex, particularly the relationship between transmission upgrades and locational generation choices. Sorting it all out inevitably leads to the demand for some form of long-run planning that encompasses the entire topic of traditional integrated resource planning. As the only operational *regional* entity, the ISO must also become the regional planner. And this planner would be driven to doing more than simply forecasting and analysis. The ISO would perforce be assuming the role of decision maker mandating investments and directing subsidies.

The six existing ISOs in the US are at different places along this path. But the trend toward increasingly complex administrative mechanisms, focused around ISO-coordinated planning, auctions and procurement mechanisms is unmistakable. In light of all this, the question that should be asked is whether the initial decision to abandon the quest for better spot price signals was correct, especially now that we have a better understanding of where the alternative paths take us.

Problems Created by Capacity Mechanisms

To understand why capacity-based mechanisms are becoming increasingly complex and administrative, it is essential to recognize that the "missing money" caused by suppressing spot prices creates two very serious problems for a market-based electricity system. First, it reduces incentives for promoting operational reliability. Second, it reduces incentives for investing in generating plant thereby threatening resource adequacy. Virtually all of the administrative complexity is designed to "fix" these two problems.

The operational problem should be obvious but it is often overlooked. If spot prices provide the signals for plants to make themselves available for dispatch, then suppressing those prices when supplies are tight and demand is high will consistently fail to signal the need for the generators to do everything they can to make their units available for dispatch. The more serious the shortage, the greater the incentive problem becomes, undermining operational reliability.

During hours of shortage when spot prices *should be* very high are precisely the times when *all* plants *should be* available in order to justify a substantial portion of the revenues need to cover their fixed costs.[28] Generators must recover these costs just to remain operational, let alone make a profit. If spot prices could rise to reflect the actual shortage conditions, generators would have strong incentives to take whatever actions were most economic to improve and maintain the reliability of their units.

If the "missing money" is made up through monthly capacity payments, generators have even less incentive to make their plants available for dispatch in shortage

[28] Actual payments to generators could be a combination of spot market and contract payments.

hours and loads have less incentive to reduce consumption during these hours.[29]

Suppressing spot prices also has an adverse effect on investment incentives. The amount of investment that will occur in response to prices will only reach a level consistent with the level of revenues provided by the markets. If the suppressed spot markets reduce the total level of expected revenues then investment in new plant will be reduced also and resource adequacy will suffer.

These causal relationships should not come as a surprise; direct analogies can easily be found in the cost-of-service regulated environment. In the past investor-owned utilities resisted building new generation to meet their service obligations when their expectations were that their future revenue requirements, would not be covered through their rates. In the late 1970s electric rates were generally held down for political reasons, causing many utilities to defer the construction of new generation projects that were projected to be needed.[30] Looking back we now know that those plants were not needed because load growth slowed in response to due to significantly higher electricity prices, but that effect was not anticipated by the industry resource planners.

III. PRICE RESPONSIVE DEMAND

A key driver of resource adequacy in an energy-only market is the existence of significant price responsive demand. But how much is "significant?" Does the MISO system have enough price responsive demand potential to meet that requirement? Assuming it does, what needs to be done to convert this potential into an actual resource? These questions are now addressed.

HOW MUCH PRICE RESPONSIVE DEMAND IS REQUIRED?

The generally accepted US industry standard for resource adequacy is a loss of load expectation (LOLE) of one day in ten years. To achieve this, a typical US power system typically requires a generation reserve margin in the range of 12 to 18 percent of peak demand. This provides operating reserve (spinning and non-spinning) and also insurance against generating unit forced outages and peak load forecasting errors.

---

[29] This incentive problem is worse if the capacity credit assigned to a unit is defined by its average availability based on forced outages over time (EFORd), as is true in Eastern ISOs. In the "LICAP" proceeding, the ISO-New England proposed to discard the familiar "Unforced Capacity" (UCAP) availability metric and replace it with an availability metric that measures whether a unit was actually producing energy or capable of providing operating reserves in real time during those hours in which the ISO experiences shortages in operating reserves. This approach in effect redefines the "capacity" product to be more like energy and operating reserves, in an attempt to make the "capacity" payment mechanisms work more like an "energy-only" market. One implication of this proposal is that a capacity payment mechanism can best solve the operational incentive problem by attempting to mimic an energy-only market with unsuppressed spot prices.

[30] This disincentive effect is known as the "Aversh-Johnson effect" and has been studied ad nauseam since introduced by two RAND Corporation economists. See Aversh H. and Johnson L.L., "Behavior of the Firm under Regulatory Constraint, " American Economic Review, Vol. 52, December 1962, pp 41-54.

Price responsive demand is defined as load the customer interrupts in response to price increases. This is load not under the direct control of the system operator, thus not providing operating reserve.[31] Since operating reserve represents about 4 to 8 percent of peak demand and will be supplied by other resources, the role of price responsive demand in directly contributing to resource adequacy is necessarily limited to reducing peak demand by about 10 percent. However, it also indirectly contributes to resource adequacy by creating the scarcity rents that make new generation profitable, thereby attracting investment in new plant.

The exact amount of price responsive demand needed to facilitate sufficient investment in new generation is determined by the following factors:

- amount of flexible load (in MW)
- price elasticity of the flexible load
- level of energy price cap
- construction and fixed operating costs of a new peaking plant
- rates of return investors require to invest in generating plant.

We intend to quantitatively investigate the effect of these factors through computer modeling of the Midwest ISO system to determine what combination will ensure that an energy-only market will produce an acceptable level of resource adequacy. The results of these efforts will be released in the near future.

Price Responsive Demand Potential In The Region

The Midwest ISO staff currently estimates the price responsive demand potential within its footprint to be in the range of 5,000 MW to 10,000 MW. This estimate is based in part on information provided by market participants to the Midwest ISO to fulfill its Module E requirements.

In addition, the above estimate of price responsive demand potential is based on the magnitudes of the large C&I loads served currently by MISO participants, combined with the price elasticities estimated in recent studies done of the real-time pricing

[31] However, price responsive demand can simultaneously participate in emergency programs controlled by the ISO, as is currently done in New York and PJM. If the customer can guarantee interruption within 10 minutes of the system operator's command it can also provide non-spinning operating reserve, as is done in New Zealand. In addition, California and international system operators have experimented with allowing customer load to provide spinning reserve as well, as long as it meets reserve specifications (*e.g.*, water pumping loads).

programs of Niagara Mohawk Power Corporation and Georgia Power Company.[32] In addition, survey data produced by EIA's most recently available Manufacturing Energy Consumption Survey (MECS), are consistent with these estimates of price responsive demand potential. Future efforts will provide our detailed analysis of price responsive demand potential within the MISO footprint.

FACILITATING PRICE RESPONSIVE DEMAND DEVELOPMENT

The incentive for retail customers to modify their electricity usage in response to price requires the Midwest ISO and state regulators to take the following actions:

- adopt scarcity pricing
- expose all generators and flexible loads to the spot market
- gradually increase the energy price cap to average VOLL for residential customers.

Scarcity Pricing

Scarcity pricing is introduced in Chapter I and discussed further in Chapter II. It reduces the frequency of involuntary curtailments in two ways. First, it encourages price responsive demand development by allowing the market price to rise above the cap on generator offer prices when generating capacity is in short supply.[33] Second, it provides incentives for generators to operate during hours of capacity when their output is needed most and for developers to add new generating capacity.

Expose Generators And Flexible Loads To The Spot Market

For generators to have the correct incentive to be available when needed, and for developers to add capacity when supply shortfalls loom, all generators should be exposed to the market-clearing spot energy price.[34] Unfortunately, price caps designed to mitigate market power violate this principle and create the "missing money" problem described earlier. For this reason the cap on market energy price should be set as high as

[32] Braithwaite, S. and O'Sheasy, M., "RTP Customer Demand Response – Empirical Evidence on How Much Can You Expect," in *Electricity Pricing in Transition (Topics in Regulatory Economics and Policy),* edited by A. Faruqui and K. Eakin, Kluwer Academic Publishers, 2002.

[33] In a LMP market some nodal prices can exceed the highest accepted generator offer even without scarcity pricing. This will occur if an increase in load forces a low-cost generator to be backed down in order to relieve a transmission constraint. The lost energy must then replaced by a higher-cost generator elsewhere on the grid. Thus, the marginal cost of serving this load increment includes not only the marginal cost of the incremental energy delivered to the load but also the increase in system cost caused by redispatch needed to relieve the constraint.

[34] One can argue that generators with bilateral contracts, or those owned by LSEs and dedicated to serving native load, need not be exposed to spot market prices. In a strict sense this is true; however, these same generators are still affected by spot market prices, albeit on a different time scale, because their alternative use before entering into a contract, or after being transferred to an unregulated subsidiary, is to sell into the spot market

practicable – ideally equal to the average VOLL for the loads that will be curtailed during hours of capacity insufficiency.[35]

Flexible loads are those that have the capability to respond to price signals on a short-term (e.g. day ahead or less) basis. These are the only loads for which exposure to real-time prices makes sense.

In theory, all loads are flexible at high enough prices, thus should also be exposed to the full spot energy prices in order for the demand curve to capture their individual valuations of electricity in any given hour. If that were so then (by definition) there would be no involuntary curtailments and (by definition) resource adequacy would be assured. However, for the foreseeable future at least, it is not cost-beneficial to meter all customers in real time (or for them to respond to those price signals even if so metered). Thus involuntary curtailments of customers lacking real-time meters will still occur on occasion and the assessment of resource adequacy will turn on the frequency and duration of such curtailments. Most of the price responsive demand response will come from a small subset of large industrial and institutional customers so the limit on meter installations is not particularly troublesome.

Exposing market participants to spot energy prices will increase their financial risk, however, most large customers will have opportunities to hedge their risk thorough forward contracting. However, residential and small C&I customers are not likely to have such opportunities available, at least not at modest cost, so it may be desirable for LSEs to serve these customers under fixed tariffs that are hedged through wholesale supply contracts or other financial instruments.

Gradually Increase The Energy Price Cap To Average VOLL

The primary reasons for capping energy prices are to mitigate generator market power and to reduce spot price volatility. As price responsive demand develops it will provide an alternate vehicle for achieving both of these objectives. This suggests a gradual relaxation of the energy price cap as price responsive demand takes over these mitigation functions. This progressive process takes advantage of the fact that as price responsive demand emerges at different price levels, suppressing market power and price volatility at each level, the price cap can be further relaxed to induce the next stage of development and price response, and so forth. In effect, the system pulls itself up by its own "bootstraps."[36]

[35] Under scarcity pricing the cap on market energy prices is conceptually different from the cap on generator offers.

[36] Numerous examples of "bootstrapping" processes are seen in the physical sciences. A thermonuclear explosion is one such example. An electric charge sets off a primer charge - which then triggers a conventional explosive - driving two masses of fissionable material together and triggering a nuclear fission explosion - which heats and compresses fusible material, triggering a thermonuclear fusion reaction.

The price cap should continue to be relaxed until it equals the average VOLL for the lowest-valued loads that are not on real-time pricing.[37] The reason for not raising the cap any further is that these customers will prefer to have their service interrupted rather than pay more for electricity than the value they place on it. In effect the spot market is acting on their behalf when it refuses to purchase energy for their use at prices above the cap.

To pursue this strategy it will be necessary to periodically assess the degree of market power remaining as price responsive demand develops so that the price cap can be relaxed as residual market power is observed to wane.

IV. MARKET POWER MITIGATION

Market power is usually an artifact of market structure rather than market design and some degree of market power will always exist in any market regardless of its design. Market power must be mitigated if reasonably competitive outcomes are to result - and allowed to produce "just and reasonable" rates.

Market power can arise if transmission constraints create pockets of captive load by limiting imports from external sources needed to restrain prices to competitive levels. In any transmission-constrained situation the potential for exercising market power is increased regardless of the market design, but in an energy-only market contrived shortages can result in spot energy prices reaching very high (shortage cost) levels. This potential for high prices provides temptation for generators to drive up prices by withholding supply.

The freedom of spot prices to reach shortage cost levels is an essential element of an energy-only market design because those prices provide the correct incentives for short-run responses from generators and flexible loads and for long-run investment in new generation and other facilities. For this reason, it is important to distinguish between real shortages and those that are contrived. The goal of market power mitigation is to preclude suppliers from profiting from contrived shortages while still allowing prices reach shortage-cost levels when there are genuine shortages.

The energy-only market design mitigates market power in several different ways. The four principal approaches are

- demand response
- limits on offer prices and must-offer rules
- reducing barriers to new entry
- contracting for power.

[37] Most likely these will be the residential customers based on the results obtained in numerous value-of-service studies conducted since the 1970s.

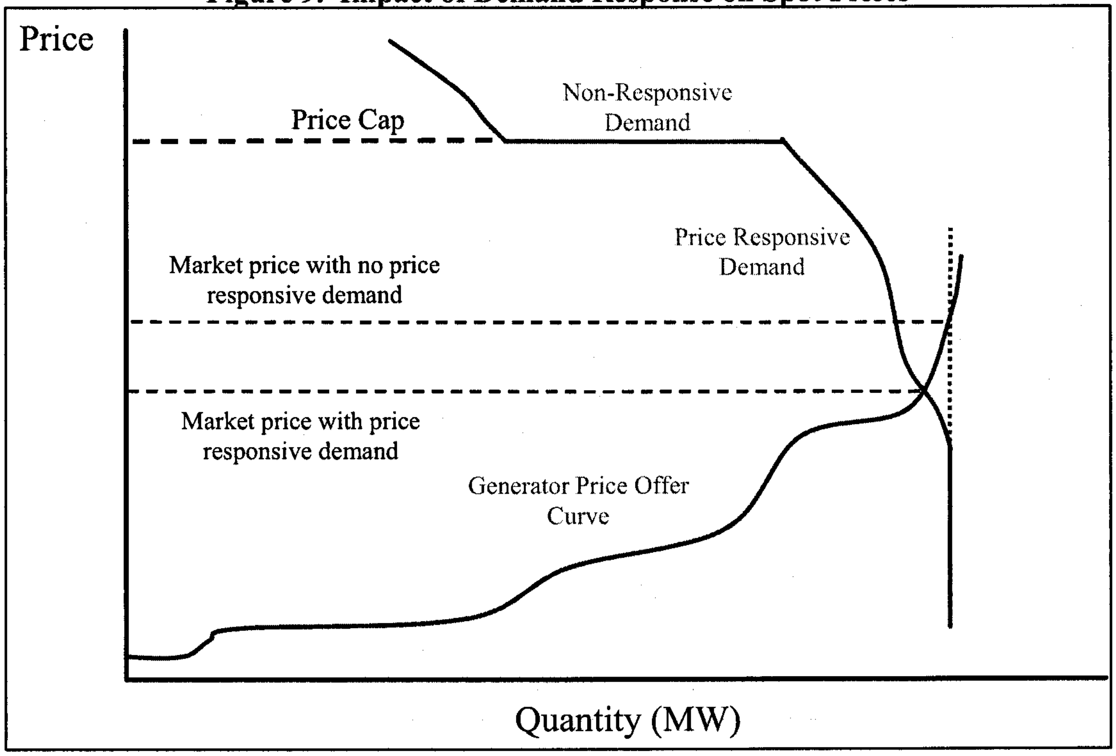
DEMAND RESPONSE

The energy-only market design allows prices to reach levels determined by customers' willingness to pay even when there is insufficient generation to meet all loads plus operating reserve requirements at lower price levels. Hence, price responsive demand is an important mechanism for limiting prices while also forestalling involuntary curtailments. As discussed in previous chapters, the price level at which involuntary curtailments occur should be the average VOLL for non-responsive customers.

This aspect of the energy-only market design stands in stark contrast to that of alternative designs that do not permit scarcity pricing. In these alternative designs prices are limited by an assumed maximum marginal cost of generation. A common misconception is that in LMP markets the spot prices at any location on the transmission grid should be determined by the marginal generator located there, and in zonal markets, by the marginal generator within the pricing zone. In contrast, the energy-only market design allows spot prices to be determined by the willingness of loads to pay prices higher than any generator offer price. In a shortage condition all generators would be operating at full output but none would be setting the spot energy price.

A relatively small amount of demand-side response can go a long way in mitigating prices when supplies are tight. As shown in Figure 9, price responsive customers will tend to reduce their loads when prices move up. The Figure shows that a small amount of demand response can produce a relatively large reduction in prices.

Figure 9. Impact of Demand Response on Spot Prices



This simple analysis clearly reveals that customers and state regulators concerned about the exercise of market power should be very interested in facilitating the development of price responsive demand.

LIMITS ON OFFER PRICES AND MUST-OFFER RULES

Supply can be withdrawn from the market either through physical withholding, whereby the generator declares some of its capacity as physically unavailable, or through economic withholding, whereby the generator submits price offers that substantially exceed the marginal cost of the capacity. In any market the temptation to withhold capacity arises from the potential that the resulting higher net revenues paid to capacity not withheld more than offset for the net revenues lost by the withheld capacity.

Mitigating Physical Withholding

In markets other than an energy-only market, the inflated spot prices resulting from physical withholding are set by the offer prices of marginal generators. In the energy-only market the spot prices could go much higher because scarcity pricing allows market prices to be set by demand-side responses up to the energy price cap set at average VOLL.

To mitigate physical withholding the market design would need to include "must-offer" rules for all units that could significantly affect market prices through physical withholding. Such rules would be reinforced with contractual incentives that discourage market power in the spot market, as discussed below.

Mitigating Economic Withholding

The market rules employed in some markets mitigate economic withholding by imposing limits on offer prices of those generating units that are needed to maintain reliable operations. Such mitigation rules would *continue* in an energy-only market, with the focus on ensuring the rules were only applied at the right times and to the right units.

Existing units needed for reliability in transmission constrained areas ("load pockets") would be the primary targets for offer price mitigation. An offer price cap would be defined for each such unit to reflect its marginal operating cost. The cap would apply whenever the plant's output was needed for reliability, such as when it is dispatched out of merit in order to relieve congestion and its plant's original offer would set the market price. These mitigated offer prices would be eligible to set the market price at each unit's respective location if the unit were on the margin.

Generating units operating at their capacity limits would not be allowed to set spot energy prices; those prices would be set by the demand side. The offer caps would serve to prevent economic withholding but prices would still reflect actual shortages in the constrained region as they should to send the correct signals. Note, however, that these units would still be paid the market price for their energy.

Because mitigated units would still receive the market energy prices, they would earn scarcity rents during periods of actual shortage, thereby avoiding the "missing money" problem. These scarcity rents would minimize the need for reliability must-run (RMR) contracts, which are used extensively in markets that are not energy-only, to keep generating units in service that might otherwise retire because of an inability to cover their fixed operating costs.

Once the costs of RMR contracts are accounted for, the total cost of the energy-only market design is not higher than those of alternate market designs. Indeed, the total cost of serving load could be lower in the energy-only market because its pricing structure provides better incentives for all generating plant to be available when most needed and for demand-side response to occur.

In any case it would be important that when units are not positioned to exercise market power they should not be subjected to these types of mitigation. Under current rules in existing markets, units in unconstrained "competitive" regions are generally not mitigated because the presence of competitors encourages availability and drives offer prices close to marginal costs. This same principle applies equally applies in an energy-only market.

REDUCING BARRIERS TO NEW ENTRY

In any market high prices attract entry of new supply. Thus, in the absence of barriers to entry, the exercise of market power sows the seeds of its own destruction. History has shown that anytime market power is exercised over extended periods there are barriers sustaining it. Reducing or eliminating such barriers is a critically important mitigation measure.

An LMP-based, energy-only market sends the right price signals to both the supply-side and the demand-side regarding where additional demand response and/or new capacity would be the most valuable. Transmission constrained areas in which LMPs are higher will attract new entry, assuming that any siting issues can be overcome.

New supply entry can come in the form of traditional utility generation, which is difficult to site in transmission constrained urban areas, or it can come in the form of smaller, strategically located distributed generation, which are generally be easier to site but which often face difficulties in getting the host utility to allow interconnection.

Distributed generation includes on-site generation owned by large customers that are exposed to real-time spot prices and it offers these customers a physical option for hedging against spot price volatility. From outside the facility these investments look like just another form of price responsive demand.[38]

[38] In Georgia Power Corp's real-time pricing (RTP) program, most of the demand response that occurs at relatively low energy prices is due to program participants running their back-up generators. See: Braithwaite, S. and O'Sheasy, M., "RTP Customer Demand Response – Empirical Evidence on How Much Can You Expect," at 32.

Streamlining the regulatory process for gaining interconnection approval, combined with net-metering rules to allow proper credit for energy produced on-site, would go far to stimulate distributed generation. In addition to mitigating market power within urban load pockets distributed generation would make the spot energy market more competitive and more efficient. A similar argument applies to small municipal systems embedded within larger utility systems.

CONTRACTING FOR POWER

The market power mitigation measures discussed above would reduce or preclude the exercise of market power by directly addressing physical or economic withholding. In contrast, contracting is a fundamentally different mitigation measure that removes the ability of a generator to profit from the withholding activity.

If a generator has a substantial portion of its capacity committed to deliver power at a fixed contract price, the revenues from the power sale are unaffected by spot prices. Furthermore, if it physically withholds any of the contracted capacity, or it suffers an outage, the generator must purchase replacement power from the spot market at the higher prices in order to meet its contract obligations. Thus, a substantially contracted generator has little or no interest in withholding supply to drive up spot market prices.

From the buyer's perspective, having its loads covered by power contracted at fixed prices also hedges the buyer (and its ultimate customers if it is an LSE) from spot price uncertainty. If spot prices occasionally reach shortage-cost levels, LSEs and their customers have strong incentives to minimize their spot market exposure. Consequently, the incentives to contract are much greater in an energy-only market than they are in other market designs that administratively cap spot prices. These enhanced incentives have positive market mitigation consequences.

Are Mandatory Contracts Needed?

The potential for spot prices to reach shortage cost levels would provide strong incentives for parties to contract voluntarily. Still, there might be reasons why state regulators would consider some form of mandatory contracting to cover customers lacking the ability to effectively hedge on their own.

Retail Choice Not Allowed

In states that do not allow retail choice regulators could provide incentives for their regulated LSEs to hedge themselves against exposure to the uncertainties of spot prices. For example, the pass-through of fuel and purchased power costs to customers could be eliminated. Or, as is currently done in California, the LSEs might be required to conduct explicit *ex ante* assessments of customer risk that are subject to prudence reviews as part of their resource planning and procurement processes. The LSEs could hedge themselves by owning their own resources or by purchasing most of their residual power requirements under fixed-price contracts. The degree and nature of utility hedging would be subject to state regulatory oversight.

Retail Choice Allowed

Regulators in states that allow retail choice would need to distinguish between two types of customers:

- those likely to choose alternative retail suppliers
- those either ineligible or unlikely to choose.

For the first customer group, it would make sense to consider whether some portion of these presumably larger customers should be subject to real-time pricing and exposure to spot prices. This exposure would provide the right incentives both for demand-side responses and for contracting to hedge exposure to spot prices. Voluntary contracting would appear sufficient for this group.

For the second customer group, the case for mandatory contracting is more compelling. Assuming an LSE has the default supply obligation, it would be expected to prudently hedge a substantial portion of its potential exposure to spot prices. It would do this through owned generation (if any) supplemented with contracts to cover any residual exposure. There are successful models for utility hedging of default supply obligations, such as the Basic Generation Service auctions sponsored by the New Jersey Board of Public Utilities, in which default supply obligations are covered by contracts acquired through competitive auctions closely monitored by independent auctioneers and state regulatory staff.

In an energy-only market it is likely that most loads would be hedged, either through contracts or through self-supplied generation, leaving a small residual amount fully exposed to the spot market. Nonetheless, this residual, together with the flexibility of loads and generators to buy from, or sell to the spot market some of their hedged energy, would provide the correct incentives for efficient behavior. Furthermore, contract prices would reflect the contracting parties' expectations of future spot prices over the contracted periods. The volatile spot energy prices would do what they are supposed to do: *provide the right incentives for short-run dispatch operations and long-run investment, while still providing most customers with relatively stable and certain electricity prices.*

V. ROLE OF STATE REGULATORS

The success of the Midwest ISO's energy-only market in ensuring resource adequacy depends not only on gaining FERC approval of its market design but also on the approval and guidance of the state regulatory commissions having jurisdiction over the load-serving entities (LSEs) within the MISO footprint. In particular, the decisions state regulators make regarding retail rate design and the supply adequacy of LSEs under its jurisdiction are critical.

State regulators are important in part because they control retail electricity rates. For an energy-only market to be successful it needs significant price responsive demand (PRD) participation. But PRD works best when at least a subset of retail customers are exposed to the hourly spot prices.[39] The state regulatory commissions control the method of billing through their approval authority over retail rate designs.

To support efficient PRD development, state regulators need to require LSEs to directly pass through their wholesale energy prices on an hourly basis to customers with real-time meters. In addition, the regulators can:

- establish minimum reserve margins to be maintained by the LSEs
- allow LSEs to count their customer PRD capacity as part of their reserve margin obligation
- require LSEs to exempt PRD capacity from charges for generating capacity.

PASS-THROUGH OF WHOLESALE ENERGY PRICES

Retail energy prices for real-time metered customers should equal the LSE's wholesale energy cost, grossed up for distribution losses. In addition, to maximize allocative efficiency the pass-through prices should reflect to the greatest extent practicable, the hourly LMPs at the retail customers' off-take nodes, rather than some load-weighted average price. State regulators would need to consider this for at least the large C&I customers that take service directly off the transmission grid.

ESTABLISH MINIMUM RESERVE MARGINS

The purpose of establishing minimum reserve margins is to guarantee an acceptable level of resource adequacy for all non-PRD customers. These reserve margins would need to be separately set for each transmission constrained sub-region within the MISO footprint and they would likely differ significantly among the sub-regions. When the energy market reaches maturity these reserve margins would seldom, if ever, be binding and the interim resource adequacy vehicle could benignly reside in the background as a "safety net." Or it could simply be dismantled.

COUNT PRD IN THE RESERVE MARGIN

PRD should be allowed to compete with generation on the proverbial "level playing field." This means that it be allowed to substitute for generation when calculating each LSE's reserve margin - at least up to the point where the remaining reserve margin, consisting of generating plant, is just sufficient to supply the entire operating reserve requirement, including provision for forced outages. The Midwest ISO's current Module E requirement reflect this 'level playing field', so this would not be

[39] Price responsive demand can work where the end users receive a lump sum fixed credit to curtail demand and the LSE, as the market participant in the Midwest ISO energy market, receives the hourly LMP reflecting the full credit for the load reduction. However, studies have suggested that persistence rates for end users in these programs diminish if the end users do not receive the full hourly LMP credit.

a change from existing practice. Considering the critical role of operating reserve it seems wise to initially err on the side of conservatism.

A megawatt of PRD need not be simply treated the same as a megawatt of generation. Although the availability of either resource is uncertain, those uncertainties are hardly equal. Further study would be needed to establish the appropriate capacity-equivalent of a MW of price responsive demand compared to a MW of peaking plant.

EXEMPT PRD FROM CAPACITY CHARGES

The surpluses of generating capacity that currently exist within the MISO footprint means that PRD may initially contribute little economic value to the system. This is because the capacity surpluses will dampen both the average spot energy prices as well as their volatilities. Thus, PRD loads will have little incentive to respond because their savings from doing so will be too small. State regulators should consider exempting PRD loads from paying any capacity charges for generation, even during this early period, because they derive no benefit from that capacity.[40] More importantly, such exemption would provide an immediate incentive for customers to implement PRD programs.[41] Until new generating capacity is needed the cost of generation capacity surpluses would need to be borne by non-PRD loads and treated as an investment in a more economically efficient future power system that will benefit all customers.

In addition to the above retail rate-related actions, state regulators can encourage PRD development in other important ways. They can:

- mandate real-time metering for all customers whose demands exceed some threshold level
- take advantage of the newly enacted EPAct – mandated requirements to fully investigate the appropriateness of real-time pricing for each customer class
- offer incentives to LSEs that promote PRD through collaborative programs
- require LSEs to hedge the price risk for small customers that cannot easily do this by themselves.

---

[40] This is not to say that the PRD loads do not benefit from the investment in, or the fixed operating costs of generation - they do. However, capacity charges are not intended to recover those costs; they are designed to reflect the marginal value of peaking capacity. PRD loads do not benefit from this capacity because they get off the system when such capacity becomes scarce. At other times PRD loads may benefit from the energy produced by peaking plants but they pay for that benefit through the spot energy market.

[41] Indeed, until spot energy prices increase significantly this could be only real incentive driving PRD development. Thus, this measure is an important part of the bootstrap strategy described earlier in the paper.

Mandate Real-Time Metering

The State of California recently invested $35 million to install real-time meters on all loads 200 KW or larger. While the concept is sound, the threshold level adopted by the Golden State may not be appropriate for the MISO states. This is an important topic for investigation by each state regulatory commission, the objective being to determine what threshold levels are cost-beneficial based on each LSE's customer price elasticities and metering costs.

State regulators have the authority to designate retail customers as either as "price responsive" or "non-price responsive," and to order the LSEs to be meter and bill the price responsive classes on an hourly basis. Initial (or pilot) programs could begin with the largest customers for whom the cost of the metering and billing systems would likely be justified. These programs could then be expanded to determine what other customer groups might be cost-beneficial candidates for real-time metering.[42]

Take Advantage Of EPAct To Fully Investigate Real-Time Pricing

Section 1252 of the recently enacted EPAct directs each state regulatory commission to conduct an investigation of the appropriateness of offering customers various types of time-based rates.[43] This same section also directs the US Department of Energy to provide technical assistance to state regulators and multi-state organizations. State regulators should take full advantage of this federal resource in their investigations of real-time pricing and PRD development.

Provide LSEs With Incentives To Promote PRD

State regulators can provide their LSEs with strong incentives to promote PRD development. Such incentives can be performance-based measures such as higher allowed rates of return or regulatory vehicles through which the LSE to share the savings with the customer. LSEs have considerable data describing the consumption patterns and behavior of their large C&I customers. Typically the LSEs have marketing representatives that establish personal relationships with the managers of these facilities. These relationships can be very valuable in influencing large C&I customers to evaluate the economic attractiveness of PRD. In addition, the LSEs could efficiently provide these customers with the know-how regarding the methods and technologies that have been successfully demonstrated elsewhere.

---

[42] Recent surveys and studies of real-time pricing using approximations of hourly spot prices suggest that the customer size for cost-effectiveness may be much lower than previously thought. See the presentations by Chuck Goldman and Roger Levy at the October 28, 2005 Restructuring Roundtable available at http://www.raabassociates.org/main/roundtable/asp?sel=65.

[43] The state regulatory commission must complete these investigations and reach decisions by January 2007.

Require LSEs To Hedge Small Customers' Price Risk

Regulators should consider requiring LSEs to contract forward for the power needed to serve small customers that lack real-time meters and cannot respond to spot energy prices. These same customers are unlikely to have the means to efficiently (if at all) hedge against electricity price risk. Although they are typically served under published tariff schedules that are fixed, sometimes for years, most are still subject to *expost* fuel and purchased power adjustments which transfer market risk directly to them with little delay. It is this risk that LSEs can hedge through forward contracting.

Although mandatory hedging on behalf of small customers is not needed for the success of an energy-only market, it does remove a potential barrier to efficient risk management while also increasing the liquidity of the forward contracts market. In addition, it reduces the potential for small customers to launch political opposition to the energy-only market design.

VI. COMPATIBILITY WITH PJM MARKET

This is an issue that will likely be resolved by the FERC. Nonetheless, all parties should be asking how the two regions would interact if the Midwest ISO adopts an EOM while PJM adopts a capacity mechanism.

Without further details about how RPM would work it is not clear that generators, given the choice of making their plants available to the Midwest ISO or PJM, would choose to give up the scarcity pricing offered by the Midwest market. Even if a generator sold its capacity in a forward PJM market, in return for monthly capacity payments, it might de-list its capacity and sell its energy into the Midwest market during actual shortages. The only way to discourage this would be for PJM to impose severe non-performance penalties on any capacity previously committed under RPM that did not offer all of its energy to PJM.

However, there is nothing about improved spot energy pricing, including the introduction of scarcity pricing, and efforts to improve demand-side response that would harm the Midwest regardless of what the FERC may ultimately decide. Moreover, it seems clear that an EOM approach in the Midwest would improve short-run reliability in this region by providing generators with the right incentives to make their plants available for dispatch when most needed.

VII. THE TRANSITION PATH

For the Midwest ISO to adopt an energy-only market a first step will be to develop a detailed roadmap specifying the tasks it must complete and the time schedule for doing so. Included in this roadmap is a menu of complementary state regulatory actions that would greatly contribute to the success of the effort. This menu includes improved retail rate designs, real-time metering and billing, and possibly mandatory contracting for energy by LSEs. Finally, the roadmap would need to address four basic issues:

- Scarcity pricing
- Operating reserve
- Market power mitigation
- Resource adequacy mechanism

Scarcity Pricing

A high priority for the Midwest ISO is to design and implement scarcity pricing with a cap on spot energy prices at the average value of lost load (VOLL) to non-price responsive customer loads. This would require reviewing available studies followed by consultations between the ISO, state regulators and market participants. Existing market rules would need to be reviewed and possibly revised to fully integrate demand-side bidding into the real-time and day-ahead markets.

As soon as these initial elements were in place, the Midwest ISO region would have the basic elements of scarcity pricing. This would provide the basic tools for the energy-only market to bring about resource adequacy. The "missing money" problem would be substantially solved, thereby encouraging investment in new supply resources. Regulatory certainty, such as a clear commitment by state regulators and market participants to the development of the energy-only market would greatly facilitate new investment. In addition, the prospect of scarcity prices would provide an incentive for buyers and sellers to enter into contracts, which in turn would provide collateral needed to support debt financing of generating plant, further spurring investment.

Operating Reserve

The ISO also needs to develop and implement software to co-optimize the economic dispatch of energy and operating reserve and to price operating reserve consistent with the pricing for energy.

Another key task is to develop a demand curve for operating reserve that is consistent with spot energy prices, as discussed in Chapter I. This demand curve plays a particularly important role in introducing some price elasticity into the spot energy market before PRD develops sufficiently to take over that function.

Eventually, it is desirable for the Midwest ISO to have bid-based regional markets for operating reserves to allow owners of resources to decide whether or not to provide operating reserve and at what price. As previously discussed, the possibility of procuring reserves through ISO-administered markets is being addressed in a parallel track through the Ancillary Services Task Force (ASTF). However, a first priority is to specify the demand curve for operating reserves that would produce energy prices consistent with any shortages in operating reserves.

Market Power Mitigation

Additional efforts focus on reviewing the adequacy of offer price caps and other rules designed to prohibit market power. The Market Monitor's review will attempt to ensure that the rules are able to identify and discourage both economic and physical withholding, and to apply mitigation selectively while not preventing prices from reflecting genuine scarcity conditions when they occurred. An assessment of the effectiveness of different amounts of price responsive demand in mitigating market power will also be conducted.

Need For A Resource Adequacy Mechanism?

A question that arises is whether the Midwest ISO should develop a capacity-based mechanism to assure resource adequacy - at least until the energy-only market is fully functional. The rationale for a dual approach to ensuring resource adequacy generally tracks the reasons often given for maintaining capacity constructs in general. Here, we discuss some of these reasons and suggest questions that parties in the region may wish to ask in deciding this issue.

Is An Interim Resource Adequacy Mechanism Needed?

The energy-only market could take some time to implement and it could take even longer for price responsive demand to reach significant levels. In the interim some mechanism would be needed to ensure that commitments to install new generation would be made in time to add the capacity when needed. Whether such a formal mechanism will be needed, and what form it should take, is best developed through extensive discussions among the stakeholders.

In any event, the resource adequacy mechanism would not need to be a capacity construct of the form adopted in the Eastern markets. A coherent proposal would need to be developed and submitted to FERC, where it would undoubtedly still face considerable opposition. Assuming some proposal could be approved, the ISO would then need to develop the software and monitoring systems to support the capacity-based mechanism. This could take longer than putting in place an energy-only market and developing the significant amount of price responsive demand needed to facilitate resource adequacy.

Is it credible to believe that such a process would take less time in the Midwest, which has no history of capacity mechanisms, than it has taken in the East? The New England LICAP process is now in its third year, is still incredibly contentious, and is still not resolved. The PJM RPM process has taken more than two years to develop and will likely take at least another year or so to litigate. Since the resource adequacy mechanism will need to be in place within the next year or two, any capacity construct appears to be infeasible. Furthermore, given the innumerable issues that would need to be resolved to develop a capacity mechanism, it seems fair to ask whether the effort would consume so much of the Midwest ISO's limited resources that the development of the energy-only market would be delayed.

Is A Capacity Requirement *Politically* Necessary To Avoid High Spot Prices?

One argument for imposing regional, or sub-regional, capacity requirements would be to reduce energy price volatility because state regulators and other public officials will not tolerate the price spikes that typically occur in an energy-only market. In response to that we need to ask whether this commonly held view is credible, given the conditions in the Midwest.

As pointed out earlier in this report, in an energy-only market most customers would not be fully exposed to *price spikes* because they would be partially or completely hedged. For example, most utilities in the Midwest still own generation and/or contract for power to serve their native load under traditional cost of service regulation. Although the utilities themselves could face volatile wholesale spot energy prices, *their customers would not see that price volatility*. This is because spot market prices would instead be averaged into the fixed retail rates customers pay, just as all supply costs are today.

For utilities that have divested their generation and/or serve under retail choice, most small and medium-size customers are still covered under state regulated rules governing default service. Those rules generally fix default retail rates over extended periods – from several months to several years – which shield customers from volatile spot prices. Furthermore, their LSEs are exposed to these spot prices only to the extent that they have chosen not to be hedged. In an energy-only market only those customers eligible for retail choice could be exposed to spot prices and most would have the ability to hedge their exposure on their own.

If so few customers would actually be exposed to the volatile spot energy market how much force does the political vulnerability argument have?

Does A Regional Capacity Requirement Make Sense For The Midwest?

There are historic reasons why the Eastern ISO markets have capacity-based requirements, and that history partly explains why those ISOs have retained capacity markets despite their obvious shortcomings. Before deregulation the Eastern markets were tight power pools that facilitated the efficient trading of economy energy among the pool members. In that environment a mechanism was needed to ensure that every pool member met its own resource adequacy requirements, because the pool's pricing rules did not penalize members that took advantage in real time of the resources built and paid for by others.[44] When bid-based markets were introduced in PJM, New York and New England, the capacity requirements were simply carried over. But after they adopted low energy price caps to mitigate market power, creating the missing money problem, the capacity requirement (or some capacity payment mechanism) became a necessity.

[44] For example, members of PJM would buy energy from each other at prices that equally "split the savings" produced by the transaction. However, if the pool ran short of energy the deficit would be physically shared among all pool members while the spot price of energy would remain limited to the pool "running rate," set by the operating unit with the highest marginal cost. PJM proudly referred to this sharing arrangement as its "three musketeers' agreement."

The history of resource adequacy in the Midwest is different from that of the East, and it differs within the Midwest. In the absence of regional power pools the need for capacity requirements to avoid free rider issues did not arise. Each utility maintained its own planning reserve requirements and in some areas no central entity allocated capacity obligations nor imposed deficiency charges for utilities that were short. The capacity requirements vary by RRO; in MAPP, for example, capacity requirements are rigorous with significant deficiency charges.

Given this very different history it is worth asking whether the usual reasons for maintaining capacity mechanisms in the East apply in the Midwest.

Is An Energy-Only Market More Complicated Than A Capacity Market?

In concept, the mechanisms of an energy-only market are straightforward and focus on getting the spot market prices to reflect the ISO's real-time economic dispatch solutions. As long as generators have a choice the price incentives they receive must support reliable operations. Consequently, getting the prices right is worth doing regardless of whether or not a mechanism also exists for compensating generators for their available capacity.

Similarly, an energy-only market works best if there is a significant amount of price-responsive demand. But there is almost universal agreement that such demand response improves any market regardless whether or not it has a capacity mechanism. However, it is also clear that price responsive demand would develop more quickly and be more effective if customers were exposed to spot energy prices that were not artificially limited by low price caps. In the energy-only market framework proposed here states would be encouraged, but not required, to expand the number of customers served under real-time tariffs to stimulate such demand-side response. Such efforts would still be worthwhile, though much less effective, if a capacity requirement were used to solve the missing money problem.

Thus, most elements of an EOM design would be worthwhile implementing whether or not the region also adopted a capacity mechanism.

If a capacity mechanism is used in lieu of appropriate spot prices, the ISO must solve three difficult problems:

- how much "missing money" must be given back to generators to achieve the region's resource adequacy goals
- how should the "missing money" be given back such that it provides generators incentives to make their plants available for real-time dispatch when most needed.
- how to induce the right amount of demand-side response - at the right times and at the right locations.

How Much To Give Back

Though generally sound in theory, the idea of giving "missing money" back to generators is a red flag for many load representatives, and the debates over this topic have been highly contentious. As long as this issue remains unresolved, it undermines all other efforts to provide regulatory certainty, and adversely impacts new plant investment.

What Availability Incentives To Provide

New England's LICAP proposal made an attempt at this, but it has been rejected at the FERC. The default is to use the UCAP mechanism proposed by PJM. But UCAP only measures availability on a time-averaged basis which does not provide the correct incentives for generator availability.

The existence of the two "missing money" problems indicate that spot energy prices are being artificially held down, which means that price responsive demand will not be strongly encouraged. This suggests the need for subsidies to correct the market pricing defect, thereby creating another potential uneconomic incentive as well as grounds for endless disagreement and opposition.

In the New York, New England and PJM reforms efforts, the "missing money" problems have exploded into an extensive and expanding list of controversial issues:[45]

- What is the reliability standard for the region? Is it a uniform reserve margin for the entire region? An LOLE standard? Different standards for different sub-regions?
- How is the standard interpreted? Is it a minimum target? An average target? A target that should not be violated more than some percentage of the time? And if the latter, how does the region decide select the data to determine the variability over time?
- How will the mechanism assure fixed cost recovery for the desired level of adequacy? What is the benchmark capacity resource, and how can we agree on its costs? Are the costs different for different sub-regions or zones?
- If the point is to restore the missing money, how does the ISO account for fixed cost recovery each generator receives from the energy and operating reserve markets? Is this contribution determined ex ante or ex post? Is it averaged, and if so, over what period? Given transmission constraints, does it vary by location? How many locations?

[45] For a lengthy description of the many design issues at only the conceptual level, see Chandley, John, *ICAP Reform Efforts in PJM and New England*, September 23, 2005, a paper prepared for the California ISO (www.caiso.com).

- How will the capacity requirement apply in different areas separated by transmission constraints? How many LICAP or RPM zones are needed? How are they defined? How can they be changed? How often will they be changed? What happens to long-run contracts when they are changed?
- How will the ISO (or FERC) define and allocate transmission rights for limited imports into constrained areas? Are the rights physical or financial? How can they be traded? What happens to these rights when the zonal boundaries change?
- What is the shape of the demand curve? Sloped or vertical? Who decides this?
- If vertical, how is the strong incentive to exercise market power dealt with?
- If the curve is sloped, what are its parameters? What is the maximum amount of capacity that warrants a payment? What is the maximum payment? What is the target for the break-even point? Who decides this? Is FERC the ultimate decision maker in deciding the curve's parameters, and thus deciding how much capacity is represented by the demand curve?
- Is capacity procured through an auction? How often? Monthly? Yearly? Four-years forward? If forward, how often must the results be updated for errors in forecasts, project delays, changed circumstances?
- Is the price determined in each auction through supplier bids? If so, how are they mitigated for market power? How is withholding prevented?
- Is the capacity price determined from the demand curve? If so, which capacity is counted? All of it? Mothballed units? De-listed units?
- How are exports handled? Can exports result in withholding? How is this checked and mitigated?
- How are imports handled?
- How is availability measured? UCAP? EFORd? Reserve shortage hours?
- If availability is measured by availability during reserve shortage hours, are there exceptions for non-availability that are not in the generator's control? Which ones? How does the ISO or its market monitor tell?

These are not hypothetical issues; they are some of the actual issues that have been or will be litigated before the FERC. Based on these observations one can only conclude that implementing a capacity construct would not simplify the market design or market operations; indeed, it would substantially complicate it.

Can Mandatory Contracting Substitute For A Capacity Mechanism?

Contracts play an important role in an EOM. Their principal purpose is to provide price certainty to buyers and sellers and hence a way to manage the risks of volatile spot prices. However, they also represent a commitment on the seller's part to have capacity available to produce the contracted energy.

Contracts for capacity can be used to circumvent the missing money problem, thereby ensuring resource adequacy. However customers will not voluntarily give generators the missing money through contracts; *they must be required to do so during the interim phase.* If contracting is voluntary, customers will be biased towards reliance on spot markets that cap prices at low levels and the region will fall short of meeting its adequacy goals.

If contracts are to be the vehicles for ensuring resource adequacy they must be mandatory. All customers (or their agents) must be required to contract for enough firm energy to cover their maximum loads or for enough capacity to cover their maximum expected loads plus a target reserve margin designed to meet the region's resource adequacy criterion. Then the ISO must ensure that the contract obligations are fairly allocated to all utilities and LSEs based on their forecasted loads, the "firmness" of their contracts (e.g., the expected forced outage rates of the plants supplying the contracted energy) obligations and load. In addition, the ISO must have some mechanism for enforcing the allocated contract obligations, such as penalties for non-performance.

Stakeholders need to consider whether mandatory contracting is really much better than other capacity mechanisms. Many of the same issues involved in designing and implementing a capacity mechanism may surface in designing and allocating the mandatory contracting obligations. If retail choice is allowed, with its potential for creating stranded costs, that will introduce further complexities.

VIII. NEXT STEPS

The Midwest ISO has initiated a work program designed to answer the primary questions surrounding an energy only market. In particular, the question of what type of market design is more likely to increase the elasticity of the real time demand curve is an important aspect of the work program. As part of this analysis, the question of what specific rules foster increased demand side participation will be evaluated. Another aspect of the work program will look at the empirical evidence related to investment in generation under different market structures.

The initial terms of reference for this work program, as well as the intermediate and final results will be presented and discussed with Stakeholders through the Working Group process. It is anticipated that the results of the analysis will not only provide guidance to the ongoing discussion on the topic but may also point out specific design elements that are necessary for a successful market.

Finally, as has been discussed in this paper, the interrelationship between the energy market and other ancillary services, in particular operating reserves, means that efforts and work programs focused on long-term capacity needs must be tightly integrated with similar efforts focused on the treatment of operating reserves within the Midwest ISO market.

Appendix: Co-Optimized Markets for Energy and Operating Reserves

While the largest and most important electricity market is the energy market, the ISO also requires ancillary services. A vital component of any energy-only market is ensuring that these providers of ancillary services are compensated appropriately. Failure in this regard can present short-term operational problems for the ISO, because it may give generators incentives not to follow operator instructions. In addition, it can undermine the market's ability to provide sufficient financial incentives to induce development of the economically efficient amount of generating capacity without the need to introduce separate revenue streams. Since prices that actually reflect the marginal value of energy or ancillary services at each location and point in time will lead to development of the economically efficient amount of generating capacity, on average, suppressing ancillary services prices below those levels will have just the same effect as would suppressing energy prices below those levels: each will reduce incentives to invest, and will generally result in less than the economically efficient amount of capacity being built.

In co-optimized markets for energy and operating reserves, the market operator simultaneously considers offers to provide energy and operating reserves submitted on behalf of each resource, and minimizes the total cost of meeting load given a set of operating reserve requirements. The price of energy then reflects the marginal cost of supplying additional energy to meet a small increase in load at each location, taking into account all of the effects of re-dispatching generation on the system to meet that increase in load at that location, including any effect of that re-dispatch on operating reserves costs. Therefore, if meeting an increment of load at that location results in an increase in operating reserves costs, that increase will be reflected in the energy price at that location. Similarly, the price of each category of operating reserves will reflect the marginal cost of supplying additional operating reserves to meet a small increase in that operating reserves requirement, taking into account all of the effects of re-dispatching generation on the system to meet that increased requirement, including any effect of that re-dispatch on energy costs. Therefore, if providing more of that category of operating reserves results in an increase in energy costs, that increase will be reflected in the price of that category of operating reserves.

As the following examples will demonstrate, settling energy and operating reserve markets using prices determined in this matter ensures that generators that are scheduled to produce a given amount of energy and provide a given amount of operating reserves do *not* have a financial incentive to disregard ISO instructions and produce more energy while providing less operating reserve, or provide more operating reserve while producing less energy. This means that the markets provide incentives for generators to follow ISO instructions, significantly reducing the need for the ISO to employ sanctions, penalties and the like to ensure that generators follow ISO instructions and provide the requested amounts of energy and operating reserve.

EXAMPLES

The following assumptions will apply to both of the examples below:

- There are three generators: A, B and C.
  - Each has 1000 MW of generating capacity.
  - Each has a minimum generation level of zero.
  - Each has a start-up offer of zero.
  - Each has submitted a single offer to produce energy using any or all of its generating capacity:
    - A has offered to produce energy for $60/MWh.
    - B has offered to produce energy for $65/MWh.
    - C has offered to produce energy for $75/MWh.
- There is only one operating reserve requirement, which states that at least 100 MW of 10-minute reserve must be maintained.
- There are no losses or transmission congestion, or requirements for other ancillary services.

Additionally, both of these examples will illustrate the calculation of prices in the real-time market. In the real-time market, there are no costs associated with making oneself available to provide operating reserves in the real-time market (as all related short-run decisions, such as whether to staff a given facility or to purchase fuel in the anticipation of being dispatched) will have been made by the time of the real-time market. Availability offers for operating reserve should therefore be zero and have been ignored in these examples.[46]

Example 1: Marginal Provider of 10-Minute Reserves is Capacity-Constrained

In addition to the assumptions above, the following additional assumptions apply to this example:

- There are 2500 MW of load.

[46] In the day-ahead market, these costs are not zero, so resources should be permitted to make availability offers, which then should be reflected in day-ahead prices. While an illustration of procedures for calculating these prices is outside the scope of this appendix, the procedures are similar to those illustrated here and produce prices that have the same properties as the prices calculated herein—namely, they ensure that each entity scheduled to provide energy and operating reserves would not be financially better off if it had been scheduled to provide more energy and less operating reserve or vice versa.

- Each generator can ramp at 4 MW/minute—meaning that each generator can provide a maximum of 4 x 10 = 40 MW of 10-minute reserve.

- The 10-minute reserves requirement is fixed, and does not depend on price.

The least-cost real-time dispatch given those assumptions is shown in Table A-1 below.

Table A-1: Least-Cost Dispatch to Meet 2500 MW of Load While Providing 100 MW of 10-Minute Reserve

Generator	Capacity (MW)	Energy Offer ($/MWh)	Energy Schedule (MW)	10-Min. Res. Schedule (MW)	Dispatch Cost ($/hr)
A	1000	60	980	20	58,800
B	1000	65	960	40	62,400
C	1000	75	560	40	42,000
Total			**2500**	**100**	**163,200**

A resource is the marginal provider of a category of operating reserves if a small increase in the amount of that category of operating reserves that must be maintained would have caused the ISO to schedule that resource to provide more of that category of operating reserves. In this case, A is the marginal provider of 10-minute reserve, as B and C are already providing as much 10-minute reserve as they can, given their ramp rates. Moreover, A is capacity-constrained, meaning that all of its capacity is used to provide energy or operating reserves, as all 1000 MW of its capacity are scheduled either to produce energy or to provide 10-minute reserve. Therefore, it faces a trade-off between providing 10-minute reserve and producing energy, as it can use each MW of its capacity to provide energy or operating reserves, but not both. If it is scheduled to provide additional 10-minute reserve, the 10-minute reserve price must recognize this trade-off, and ensure that A does not have an incentive to ignore ISO instructions.

If the 10-minute reserve requirement were to increase by a small amount, to 101 MW, the dispatch would change as shown in Table A-1A below.

Table A-1A: Least-Cost Dispatch to Meet 2500 MW of Load While Providing 101 MW of 10-Minute Reserve

Generator	Capacity (MW)	Energy Offer ($/MWh)	Energy Schedule (MW)	10-Min. Res. Schedule (MW)	Dispatch Cost ($/hr)
A	1000	60	979	21	58,740
B	1000	65	960	40	62,400
C	1000	75	561	40	42,075
Total			**2500**	**101**	**163,215**

The additional 10-minute reserve must come from A, but that reduces the amount of energy that A can provide, since A is capacity-constrained. This energy would have to be made up by C, which is the least expensive unit with capacity available. Therefore, the increase in the total cost of the dispatch resulting from an incremental change in the

10-minute reserve requirement consists of the increase in energy costs resulting from shifting one MWh of output from A to C. This is \$75 – \$60 = \$15, so the price of 10-minute reserves is \$15/MW.

The energy price in this example is simply C's \$75/MWh bid, since C is the marginal provider of energy (and is not capacity-constrained). We can therefore see that A would not be made better off by producing more energy, since it realizes a margin of \$75 – \$60 = \$15/hour on each MW used to produce energy, and it is paid \$15/MW for each MW used to provide 10-minute reserve. We can also see that B and C actually *prefer* providing 10-minute reserve to generating energy, since their margins on generating energy are less than \$15/MWh. Consequently, each would like to provide the maximum amount of operating reserve it is capable of providing, and each has been instructed by the ISO to do so, consistent with its wishes.

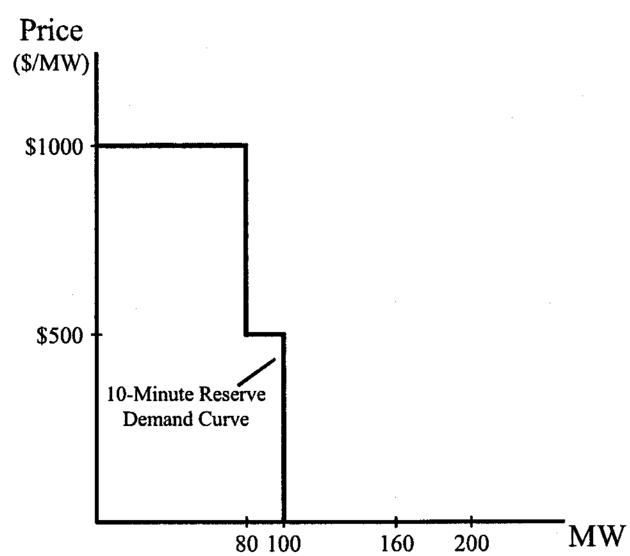
Example 2: Marginal Provider of Energy is Capacity-Constrained

In place of the assumptions above, the following additional assumptions apply to this example:

- There are 2910 MW of load.
- Each generator can ramp at 6 MW/minute—meaning that each generator can provide a maximum of 6 x 10 = 60 MW of 10-minute reserve.

Additionally, assume that the 10-minute reserves requirement is set using a demand curve, which is set at \$1000/MW for the first 80 MW of reserve meeting the requirement and \$500/MWh for the last 20 MW of reserve meeting the requirement, as shown in Figure A-1. In other words, the 10-minute reserve should normally be 100 MW, but it may be reduced to as little as 80 MW if the price of 10-minute reserve reaches \$500/MW, and it may be reduced to zero if the price of 10-minute reserve reaches \$1000/MW.

Figure A-1: Demand Curve for 10-Minute Reserves



One can consider the reserve demand curve (RDC) to be a "provider" of 10-minute reserve. Of course, it does not actually provide 10-minute reserve, but reducing the requirement using the RDC has the same effect on prices and schedules as an actual generator that is offering to provide up to 20 MW of 10-minute reserve at $500/MW and up to another 80 MW of 10-minute reserve at $1000/MW. For example, if the ISO only procures 90 MW of 10-minute reserve because 10-minute reserve prices have risen to $500/MW, then the actual generators would be scheduled to provide 90 MW of 10-minute reserve, just as they would have been scheduled to provide 90 MW of 10-minute reserve if an actual generator offering to provide 10 MW of reserve at $500/MW had suddenly appeared.

The least-cost real-time dispatch, given those assumptions, is shown in Table A-2 below.

Table A-2: Least-Cost Dispatch to Meet 2910 MW of Load with 10-Minute Reserve Requirement Determined Using Demand Curve

Generator	Capacity (MW)	Energy Offer ($/MWh)	10-Min. Reserve Cost ($/MWh)	Energy Schedule (MW)	10-Min. Res. Schedule (MW)	Dispatch Cost ($/hr)
A	1000	60	0	1000	0	60,000
B	1000	65	0	970	30	63,050
C	1000	75	0	940	60	70,500
RDC	20		500		10	5,000
RDC	80		1000		0	-
Total				**2910**	**100**	**198,550**

B is the marginal provider of energy in this example, since A, the unit with the lowest energy offer, is already producing as much energy as it can. B is also capacity-constrained, since all 1000 MW of its capacity are scheduled either to produce energy or to provide 10-minute reserve. Therefore, if the 10-minute reserve price is high (as it will be if the RDC is being used to reduce the 10-minute reserve requirement), and B is scheduled to provide additional energy in response to a small increase in load, the energy price must be calculated in a manner than ensures that B has an incentive to produce energy instead of providing 10-minute reserve, if B is to provide energy voluntarily.

If load were to increase by a small amount, to 2911 MW, the dispatch would change as shown in Table A-2A below.

Table A-2A: Least-Cost Dispatch to Meet 2911 MW of Load with 10-Minute Reserve Requirement Determined Using Demand Curve

Generator	Capacity (MW)	Energy Offer ($/MWh)	10-Min. Reserve Cost ($/MWh)	Energy Schedule (MW)	10-Min. Res. Schedule (MW)	Dispatch Cost ($/hr)
A	1000	60	0	1000	0	60,000
B	1000	65	0	971	29	63,115
C	1000	75	0	940	60	70,500
RDC	20		500		11	5,500
RDC	80		1000		0	-
Total				**2911**	**100**	**199,115**

In order to meet an incremental MWh of load at the lowest cost, the system operator would dispatch B to generate an additional MWh of energy while reducing the 10-minute reserve by an additional 1 MW. The change in the total cost of the dispatch that results from the need to meet an increment of load consists of A's energy offer plus the RDC price, since the RDC is "providing" an additional MW of 10-minute reserve. This is \$65 + \$500 = \$565, so the price of energy is \$565/MWh.

The 10-minute reserve price in this example is simply the \$500/MWh price on the lower step of the RDC, since that portion of the RDC is the marginal provider of 10-minute reserve. B would not be made better off by producing more energy, since it realizes a margin of \$565 – \$65 = \$500/hour on each MW used either to produce energy, which is equal to the \$500/MWh it is paid to provide 10-minute reserve. A prefers generating energy to providing 10-minute reserve, since it realizes a \$565 – \$60 = \$505/MWh margin when it generates energy, which exceeds the 10-minute reserve price, so it will be willing to follow the ISO's instructions to produce energy with all of its capacity, despite the high 10-minute reserve prices. C prefers providing operating reserve to generating energy, since the margin it realizes when it generates energy is only \$565 – \$75 = \$490/MWh, so it will be willing to provide the maximum amount of operating reserve it is capable of providing, just as the ISO has asked it to do.