



Andrew L. Ott
Vice President, Markets

955 Jefferson Avenue
Norristown, PA 19403
Office: 610-666-4267
Fax: 610-666-4281
ott@pjm.com

**PJM's Reliability Pricing Mechanism:
(Why It's Needed and How It Works)**

The attached paper explains in basic terms PJM's long-term resource adequacy program which is called the Reliability Pricing Model or RPM. The paper discusses why RPM is needed now to assure the adequate supply of energy in future years and how it is designed to provide the correct market based incentives to encourage the right mix of resources. It also relates how some of the basic provisions of RPM have been developed / considered by other regional transmission organizations / independent system operators. It is written for individuals who have some knowledge of long term resource adequacy proposals.

The paper, which was commissioned by PJM Interconnection and authored by Mr. John Chandley, a principle in the electricity market design group of LECG, is part of an effort by PJM to continue to explain how PJM's electricity market functions and helps PJM to keep the lights on today and tomorrow.

Sincerely,

Andrew L. Ott
Vice President, Markets
March 2008

PJM's Reliability Pricing Mechanism: (Why It's Needed and How It Works)

John Chandley, LECG, LLC.¹
March 2008

Introduction

PJM's Reliability Pricing Model (RPM) is a long term resource adequacy mechanism designed to solve one of the most difficult challenges in the design of wholesale electricity markets. The challenge is whether "market-based" prices in those wholesale markets can provide the right incentives to support investments in the appropriate amount and types of electricity supply and demand-side resources. Closely related is whether market prices encourage the right mix of resources to be available in those hours when most needed, as when there are not enough resources to meet energy plus operating reserve requirements.¹ These questions thus go to the heart of how successful electricity markets can be.

Addressing the challenge of getting the incentives right is the principal motive for RPM and drove its design. If the wholesale market rules properly set these incentives, investors will support the development of enough generation and demand-side resources, and the investments will include an appropriate mix of resources with the types of operating characteristics that allow the electricity system to be operated reliably and efficiently. If the incentives are wrong, investors may not develop enough resources, and even if enough physical capacity is built, the resources may not be available when needed or possess sufficient operating flexibility. Getting the wholesale price incentives right is thus fundamental to keeping the lights on at the lowest cost.

Given the complexity of electricity operations, the incentive issues would arise in any event, but they are most easily understood in the context of wholesale electricity spot markets administered by Regional Transmission Organizations (RTOs), such as PJM.²

¹ Electricity is a highly complex bundle of products, but the basic product consumed by end-use customers is "energy," measured in kilowatts hours on their monthly bills, and generated by power plants in megawatt hours. Operating reserves consist of additional generating capability held on standby to replace operating plants that experience sudden, unexpected outages. Operating reserves must begin producing energy quickly to rebalance supply and demand, thus avoiding blackouts or other reliability concerns.

² RTOs, which operate at the wholesale level, typically operate both a day-ahead spot market and a real-time spot market. Parties can buy and sell energy in the day-ahead market, locking in day-ahead prices for the energy they expect to produce or consume the next day; or they can buy and sell energy on an hourly or even shorter basis in the real-time market, with prices set in "real time" – that is, at the moment the energy is

RTOs are regional organizations that operate regional power pools.³ RTOs operate wholesale spot markets as an integral part of the physical dispatch of generation, the process RTO system operators use to keep supply and demand balanced while maintaining adequate voltage and keeping flows across each transmission element within safe operating limits.

The problem is that RTO wholesale spot market prices for the two most basic electricity products, energy and operating reserves, do not fully reflect their value during periods of shortages.⁴ Every RTO has one or more rules that limit spot prices and keep them from reaching shortage-cost levels. The overall effect is to reduce the revenues generators receive in hours of operating reserve shortages, with a resulting reduction in the financial incentives for long-run investments, short-run availability and operational flexibility. Similarly, because the rules limit prices during shortages, there is less investment in demand-side response capability, coupled with reduced incentives for demand-side responses precisely when they would be most valuable in lowering prices and relieving shortages.

The shortage pricing problem has been difficult to solve, because there are as yet insufficient mechanisms for end use electricity consumers to signal the value they place on being served versus being curtailed. Historically, this meant the demand for

physically generated and consumed. Other wholesale markets for long- and short-term contracts use the RTO spot markets for balancing and other support.

³ Power pools have been around almost since the beginning of electricity power systems – PJM began in 1927 -- and are common throughout the industry, in all countries. Using the transmission system to move power, virtually every electric utility “pools” its generators and coordinates their operations to serve the collective demands of the utility’s customers, rather than having specific generators serve only specific customers. Utility pools typically serve only their customers; RTO power pools tend to consolidate the dispatch and transmission operations of multiple utilities across a larger region, but the “pooling” concept is essentially the same: coordinated dispatch of all generators in the “pool” to serve all the demands (loads) in the aggregate system. Because RTOs operate transmission systems that cross state boundaries and thus move power in interstate commerce, their operations are subject to oversight by the Federal Energy Regulatory Commission, while most of the retail rates for end use customers of the individual utilities within the RTO pool are regulated exclusively by state utility commissions.

⁴ “Shortage” does not necessarily mean the extremely rare conditions in which there are insufficient supplies to keep the lights on. Far more often, it refers to conditions in which supply and demand-side response resources are less than the total demand for energy *plus operating reserves* (plants held in reserve in case plant outages occur), so that the system operates with less than the desired level of reserves even though all loads are being served with energy and no involuntary curtailments have occurred.

reliability, whether for planning reserve margins⁵ or operating reserves, has been administratively determined by regional reliability organizations or the RTOs themselves. Further, most RTOs do not yet have all the needed markets for pricing the value of operating reserves or integrating their pricing with the pricing for energy. PJM has markets for regulation and spinning reserves but is still developing markets for other types of reserves. In addition, occasional prices at very high shortage levels could raise political concerns, even though most end use consumers receive service under fixed retail rates and would not directly face such price spikes.

Whatever the reasons, RTO rules tend to limit wholesale spot prices during hours of shortage. The lost revenue effect is sometimes called the “missing money” problem, in which wholesale market revenues prove insufficient to support the level and types of investments deemed necessary to support reliability objectives, while administrative price caps during hours of shortage reduce the incentives needed to encourage appropriate levels of supply and demand-side responses when they are most needed and valuable.

Like other capacity-based mechanisms, PJM’s RPM seeks to address the missing money problem by providing payments to reward demand-side resources and pay generators the money they would otherwise receive if wholesale market prices for energy and operating reserves properly reflected the value during shortages – that is, the full value when the system is short on operating reserves. In other words, RPM’s “capacity payments” are not designed to give generators “windfalls” nor do they subsidize investments at capacity levels that would not be approved under cost-of-service regulation. Rather, RPM takes as a given whatever reliability targets the region adopts (such as a 15 percent planning reserve margin), and then seeks to provide the missing revenues that would be needed to support the corresponding level of investment.

RPM’s more advanced features recognize that payments must also be structured to provide the right incentives at the right geographic locations, depending on whether a sub-region of the transmission grid has limited ability to import electricity. Transmission limits into an area may restrict access to lower cost supply, forcing wholesale prices there to be higher, while prices in unconstrained areas are lower. Hence, a uniform payment to all generators across PJM, regardless of location, would provide the wrong incentives – either too high or too low --to resources at each location. The problem was becoming particularly severe in transmission-limited areas, where new investments were needed but were not being pursued.⁶ The RPM solution allows capacity payments to differ in

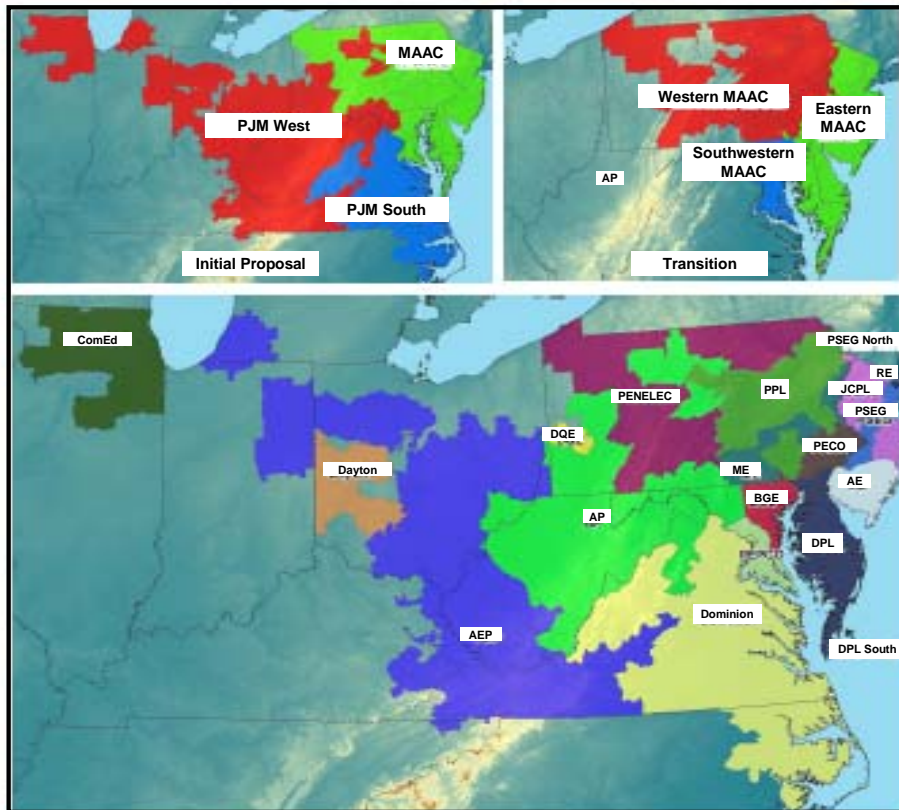
⁵ Planning reserves usually refers to additional capacity available to a region above the amount needed to meet expected peak demand. These extra reserves reduce the likelihood that unexpectedly high demands or generation outages will result in insufficient supply to meet demand. Traditionally, utilities built enough capacity to provide something like 15 percent planning reserve margins.

⁶ With uniform payments, capacity payments in constrained areas tended to be too low to sustain plants needed for local reliability. To encourage more investment and/or prevent premature retirements of local units needed for reliability, RTOs made additional

different sub-regions of PJM, with higher payments made in locations with limited transmission and lower payments elsewhere.

As shown in Figure 1, RPM divides the PJM footprint into different locational deliverability areas (LDAs) and conducts simultaneous auctions for each LDA; when transmission limits occur, this allows prices in transmission limited LDAs to rise while prices are lower in other areas with more transmission (or surplus capacity).

Figure 1
RPM Divides PJM Into Locational Deliverability Areas
New LDAs Will Phase in Over 3 Years



Another important objective of RPM is to condition capacity payments on performance, so as to improve the incentives for resource to be available in those hours most needed and with the right operating flexibility. Merely paying generators for installed capacity – or “iron in the ground” -- provides little incentive for generators to make their units available in those hours, nor does it encourage investors to build plants with quick-start,

payments to “reliability-must-run” (RMR) plants, based on cost-of-service principles. As LDAs are phased in over three years, RPM’s locational pricing should gradually reduce the need for RMR contracts.

ramping and dispatch flexibility. These are valuable features in keeping the lights on when the system is stressed.

RPM addresses part of this problem by conditioning the amount of capacity payment a resource provider receives based not only on the resource's general availability but also on its specific availability in those hours in which the PJM system is short on operating reserves. Since reserve shortage hours are those in which energy and operating reserve prices are most likely to be limited by RTO wholesale pricing rules, the effect is to focus more of the payments on precisely those hours in which the "missing money" problem arises, thus connecting with the underlying rationale for RPM.

Because RPM differs from PJM's previous capacity payment mechanism, the transition to RPM has created concerns, and this paper seeks to explain how the concerns arise and why they are either premature or misplaced. (A forthcoming companion paper will address this in more detail.) For example, one misplaced concern is that with the payments for capacity under RPM, wholesale electricity markets will result in total costs (and eventually retail rates) that are higher than what would occur in a full cost-of-service regulatory regime. This is not correct. As this paper explains in later sections, the final prices realized under RPM are limited by cost-of-service principles, so that over time, total market revenues from PJM's energy, operating reserve and capacity markets (RPM) cannot exceed the total revenue requirements that would be determined under well functioning cost-of-service regulation.

To be sure, capacity payments in the initial phases of RPM should be higher than capacity payments were under the installed capacity (ICAP) markets that preceded RPM. This effect is transitional and was expected. The primary reason for this increase is that RPM's new downward-sloping demand curves more accurately reflect the value of capacity during "surplus" conditions. That value is not zero. In contrast, the previous ICAP vertical demand curve implied that even a small amount of excess capacity had little or no value, while small shortages of capacity could result in extremely high capacity prices, a feature that created both extreme volatility in capacity prices and strong temptations to exercise market power by withholding capacity from the ICAP auctions. RPM's sloped curves flatten out these extremes and smooth out payments over time, allowing for greater price stability.

In the sections that follow, this paper describes the background that led to the reforms embodied in RPM and further explains RPM's key features. It illustrates the incentive problems associated with the "missing money" problem and traces how PJM and other RTOs developed solutions that eventually evolved into RPM. Along the way, the paper addresses various concerns that have been raised about RPM.

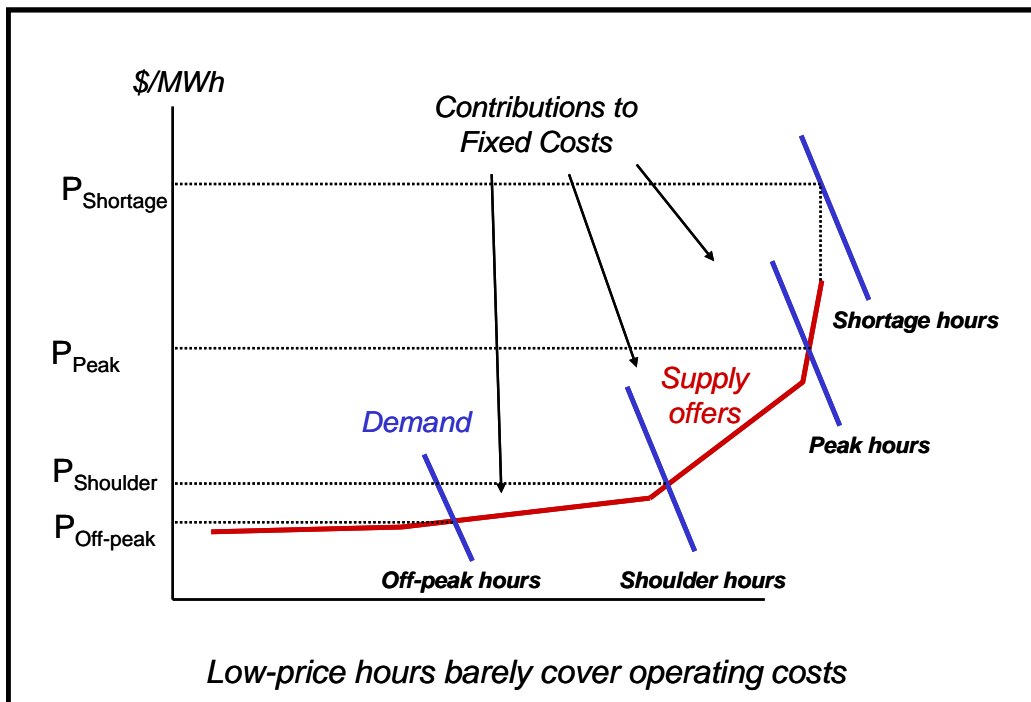
Origins of the Missing Money Problem

The need for a mechanism like RPM is driven by what analysts call the “missing money problem.” These terms describe what happens when an RTO’s spot pricing rules limit the prices paid for wholesale spot energy and operating reserves below market-clearing levels during periods when the system operator is short of operating reserves.

This shortage condition need not be so extreme as to require rotating blackouts, an emergency condition that, given today’s 15 percent or higher planning reserve margins, is not likely to occur more than once in a decade on average. More often the shortage occurs when PJM (or any system operator) does not have sufficient resources available for dispatch to meet the combined demand for energy *plus the requirement for operating reserves*. Operating reserve shortages can therefore occur when there are still enough resources to meet energy demand, so that no customers are involuntarily curtailed, but the system operators do not have enough resources left over to meet their desired level of operating reserves.

To see the effect of the absence of scarcity pricing, it is first necessary to show how wholesale spot prices allow generators to recover their fixed costs. Figure 2 provides a highly simplified illustration of how PJM’s spot energy prices might change over the day.

Figure 2
Generators Depend on the Highest Price hours to Recover Most of Their Fixed Costs



As shown in Figure 2, there is a set of generation units available for dispatch and to meet operating reserve requirements, represented by the supply curve. The supply offers of available generating plants are arranged in merit order of their increasing operating offers (assumed to represent their variable operating or opportunity costs), so that the plants with the lowest operating costs (those represented by the left side of the curve) will be dispatched first, followed by plants with increasingly higher operating costs (on the right side). Several demand curves are shown for different periods of the day; the demand curve shifts to the right during the day as demand for energy plus operating reserves rises, before falling later.

During off-peak periods, only the plants with the lowest operating costs (as indicated by their offers) are needed, so the market clears at fairly low prices ($P_{\text{Off-peak}}$). At these prices, the cheapest plants to operate recover their operating costs; they recover little or nothing towards their fixed/capital costs. As demand increases during the day (e.g., to shoulder hours) somewhat more expensive plants are needed, raising the clearing price. The corresponding prices cover the operating (as-bid) costs of the dispatched plants, and they also begin to contribute some revenues to cover the fixed/capital costs of the plants with lower operating cost. During peak-period hours, demand is much higher, requiring more expensive plants to operate and pushing prices up to relatively high levels. At these peak-hour prices (P_{Peak}), even more plants are receiving some contribution to their fixed/capital costs.

This sequence of changing prices is typical; it occurs every day, and tends to be moderated during weekends. The magnitudes also vary from week to week and over the seasons, with much higher peak-period prices typically occurring during the summer months when even the most expensive plants to operate are needed to meet peak energy demand or provide operating reserves. All but the most expensive plants recover significant portions of their fixed/capital costs during these peak hours.

The more interesting point depicted by Figure 2 is what happens on those rare occasions when the total demand for energy plus operating reserves exceeds the total supply. This might occur only a few hours each year, such as during extreme, extended heat waves and/or when a higher than expected number of large generating plants or transmission lines are out of service. In some years, this condition might not occur at all; in other years, it may happen several times, for an hour or more each time. It is during these hours when shortages exist and “shortage-cost pricing” should kick in, but current RTO rules do not yet fully provide for shortage-cost pricing; prices are kept at lower levels.

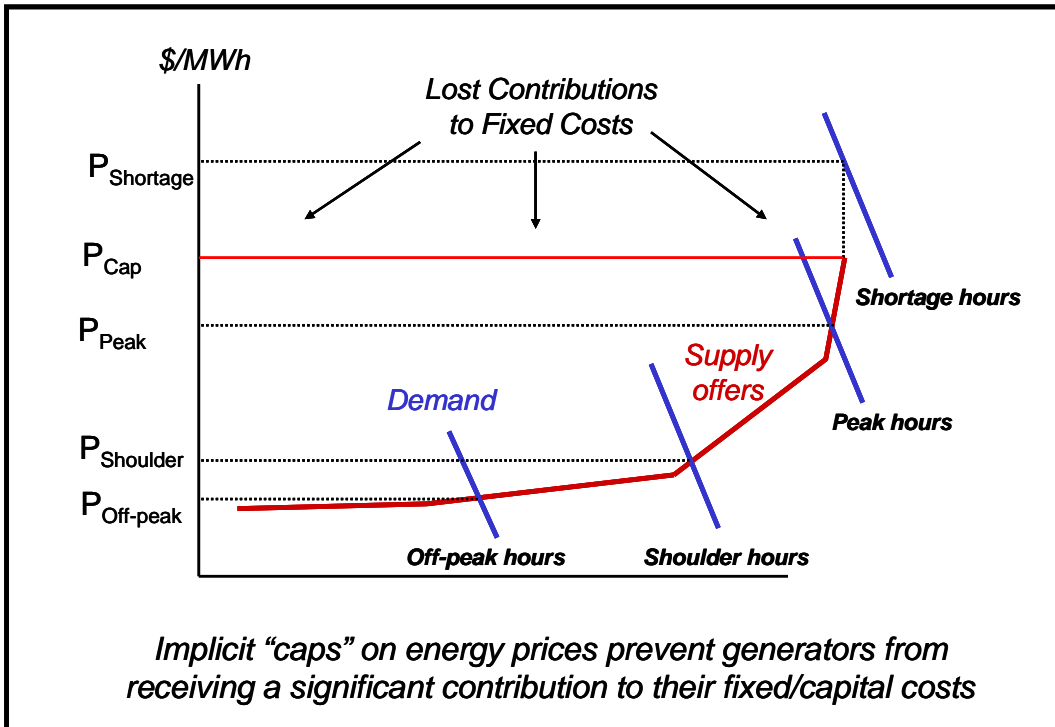
Under shortage-cost pricing, PJM would specify demand curves for energy plus operating reserves, and when supplies are insufficient to meet both, the demand curves would set a clearing price above the operating cost (or offer price) of the marginal plants – usually (but not always) the last plants available on the supply curve. Figure 2 provides a simplified illustration of this condition, showing a very high “shortage price” (P_{shortage}) clearing the market during a shortage of operating reserves.

Contributions to fixed/capital costs would be particularly high during these rare shortage hours, as Figure 2 shows. Over time, every plant in the system needs at least some shortage hours with shortage-cost prices to ensure they recover their full fixed/capital costs. If prices were never allowed to reach shortage-cost levels, the total market revenues would be insufficient to cover the set of fixed/capital costs associated with this mix of generation. Over time, investment would decline, existing plants would be retired and not replaced, and the total capacity would shrink to match the available revenues.

Like most RTOs, PJM does not yet have a complete shortage-cost pricing mechanism for its wholesale spot markets. Nor does it have a complete set of operating reserve markets with reserve prices integrated with energy prices. However, PJM does allow generators to raise their offer prices (and not be mitigated for market power) under conditions of operating reserve shortages. The rule thus allows higher supply offers to raise prices towards shortage levels; the rule is thus an attempt to mimic part of the effect of shortage pricing. Even when this occurs, however, PJM spot energy prices are unlikely to rise as high as they probably would under full shortage-cost pricing. In addition, like other RTOs, PJM limits generator offer prices to \$1000/MWh, and some plants may have their offers further reduced by PJM under market-power mitigation rules.

Figure 3 illustrates this condition as an implicit cap on energy prices (P_{Cap}). The combined effect of these rules prevents wholesale spot energy prices from ever reaching shortage-cost levels.

Figure 3
Implicit Price Caps Create the Missing Money Problem



As Figure 3 shows, all generators, not just peaking plants, are affected by the energy price limits, because every generator is denied the full price it would otherwise receive during shortages of operating reserves. The missing money problem affects the entire fleet, not just peakers. Because the total revenues provided by market prices are lower, the incentives for and level of investment will also be lower over the long run. Thus the fleet of plants assumed in Figures 2 and 3 cannot be sustained over time; as units retire, fewer plants will be built to replace them. If an RTO decides to use capacity payments as the means to replace this missing money, such payments must therefore be made to all generators, not merely to peaking plants.

The adverse effects are not limited to reduced long-run investments. The implicit price cap depicted in Figure 3 only kicks in during periods in which there are serious supply shortages, at least serious enough to reduce the amount of operating reserves available to back up the system in case one or more operating plants suddenly experience an unplanned outage. Although all energy demand may still be served, the shortage of operating reserves places the system at risk of potentially catastrophic collapse; if the shortage condition worsened, and a contingency occurred, it would force the PJM system operators to shed some load to protect the total system from widespread blackouts.

Operating reserve shortage hours are exactly the time when the system would benefit from the strongest possible incentives for price responsive customers to voluntarily reduce demand and for generators to make extraordinary efforts to make any remaining capacity available for dispatch. With such efforts, PJM could improve reliability and reduce the threat of system collapse. But the implicit cap on prices caused by the absence of shortage-cost pricing blunts the incentives that both supply and demand-side providers receive. With suppressed prices, it may not be worthwhile for as many suppliers to take those extraordinary steps to become available, or for price-sensitive customers⁷ to reduce or shift their demands to less critical hours.

The “missing money” problem is thus a problem of missing incentives, both long-run incentives for investment and short-run incentives to encourage availability and operating flexibility. Any capacity-based mechanism, such as PJM’s RPM, must therefore address all of the incentive problems created by the absence of shortage-cost pricing. And it must do so by channeling payments in ways that not only provide the missing money, but equally important, pay that money for availability and performance in the hours when resources are most needed.

⁷ Price sensitive customers include utilities or load-serving entities responsible for demand response programs; shortage-cost pricing would trigger these response programs. In some states, the largest end-use customers have default retail rate designs that track wholesale spot prices, so that during shortage-cost periods, price-sensitive end-use customers could also respond directly to high prices by curtailing and/or shifting demand. The benefits of such responses in mitigating prices and shortages (and reducing the need for more peaking plants) are principal reasons why several states are moving to adopt such rate designs.

Why RPM Needed a Different Demand Curve: Pre-RPM Market Experience

Since their inception, all the Eastern power pools that later became RTOs or Independent System Operators (ISOs) – PJM, New York ISO and New England ISO -- have required load serving entities (LSEs) to meet their share of a regional capacity obligation. In a power pool, generating units from all pool members are centrally dispatched by the pool (now RTO/ ISO) and the entire region relies on the collective supplies from all plants being dispatched at any given moment. The common dispatch also means any sub-region or member utility can, at least for short periods, rely on other members' power plants and transmission to help meet its loads.

Each pool had a regional capacity obligation, which was defined as a planning reserve margin sufficient to meet an industry-wide reliability standard.⁸ Before the pools became ISOs (or RTOs) with open dispatches⁹ and associated spot markets, the pool members recognized that each member had to accept an obligation to meet its share of the regional capacity obligation, corresponding to that member's share of the pool's peak demand plus planning reserve margin. Otherwise, members who were "short" on capacity to cover their own loads could unfairly lean on other members and avoid paying the fixed/capital costs of power plants needed to meet the required reserve margins. This individual member capacity obligation continued when the power pools became ISOs, and it was then adapted to "load-serving entities" (LSEs) when states began to permit retail choice among LSEs other than local utilities.

When retail choice began, new LSEs were not always utilities and did not necessarily own generation, but in the emerging electricity markets, it made no sense to require every competitive retail supplier to build its own generating plants. As the ISOs developed spot energy and operating reserve markets, it became clear that the LSE capacity obligations could more easily be satisfied if LSEs could purchase "capacity" – i.e., meet their capacity obligation -- from monthly auctions administered by the ISO. Utilities and other generation companies could in turn sell capacity through these same auctions.

The monthly auctions made it appear that "capacity" was being traded as an identifiable physical product. But from a financial perspective, these Installed Capacity (ICAP) markets were (1) a means to require every LSE to pay its share of the fixed/capital costs

⁸ A common US "standard" is to build enough generation so that the chances of having insufficient supply to meet peak demand will not occur more than once, or one day, every ten years, on average. A 1-day in 10-year Loss of Load Expectation (LOLE) translates to approximately 15 to 20 percent planning reserve margins above expected peak demand. The percent varies depending on the expected outage rates of the generators and transmission lines available to the system.

⁹ Any generator, regardless of ownership, can offer its capacity to PJM for centralized pool dispatch.

of ICAP and (2) a means to pay ICAP providers for the fixed/capital costs that were not recovered from profits in the energy/operating reserve markets.

In the ISOs' monthly ICAP auctions, the price offers from generation suppliers defined a supply curve. The "demand" curve, however, did not represent the buyers' willingness to pay for capacity; rather it represented a demand fixed by administration rule: a vertical line corresponding to the amount of installed capacity needed to meet the region's target reserve margin, typically about 15 percent of projected peak demand. The intersection of capacity supply and the vertical administrative demand curve would then define a monthly price for ICAP.

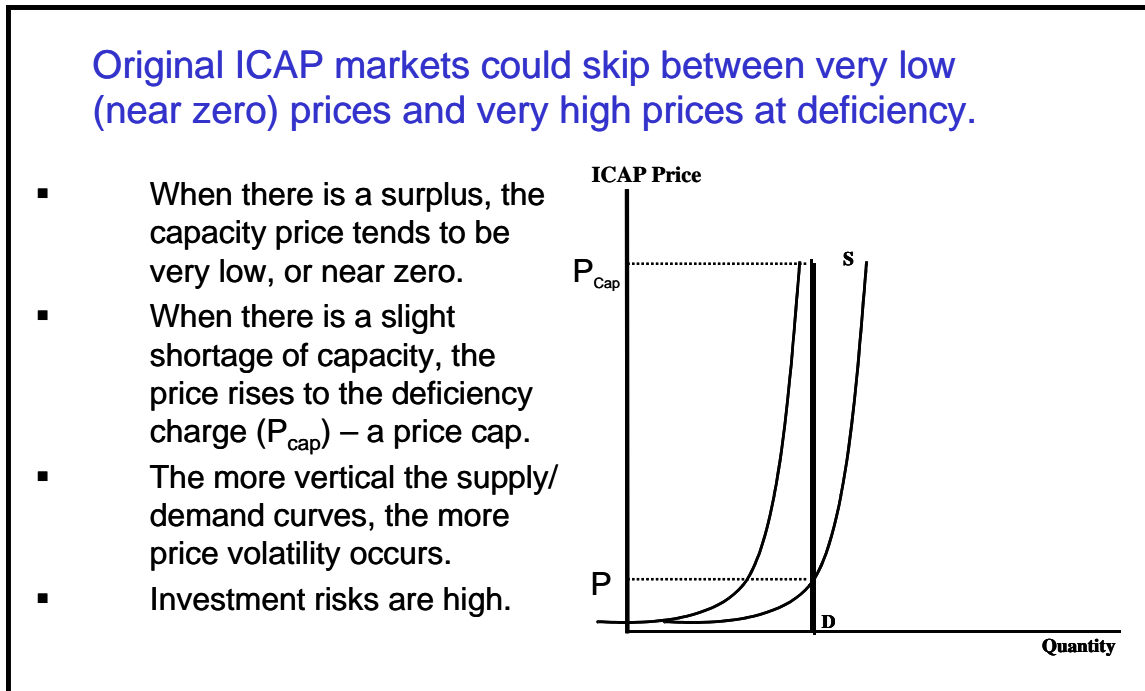
Even with ICAP payments, PJM markets failed for several years to provide sufficient revenues to cover the full fixed costs of the level of resources needed to meet PJM's regional planning reserve requirements. In the *2006 State of the Market Report* the PJM market monitor analyzed whether the fixed costs of new peaking, mid-merit and coal-fired base load were covered by the prices received by generators from the PJM energy, capacity and ancillary services markets. The market monitor concluded that:

"Analysis of 2006 net revenue, including both the Day-Ahead and Real-Time Energy Market, indicates that the fixed costs of new peaking, midmerit and coal-fired baseload were not fully covered. During the eight year 1999 to 2006, the data lead to the conclusion that net revenues were less than the fixed costs of generation and that this shortfall resulted both from lower, less volatile energy market prices and lower capacity credit market prices in the last several years."¹⁰

Figures 4 and 5 illustrate two additional problems the ISOs encountered with this earlier approach, both of which stemmed from using a fixed vertical demand curve. The first problem was that prices could be very volatile from month to month. If the region had a slight surplus of capacity (e.g., slightly more than 15 percent reserve margins), then the price could be very low. Eastern ISOs often experienced months when the monthly capacity price was close to zero. In other months, a slight deficit in supplies would push capacity prices to a very high price cap, which might be some multiple of the capital costs of a new combustion turbine generating facility.

¹⁰ PJM, *2006 State of the Market Report*, vol 1, p. 15.

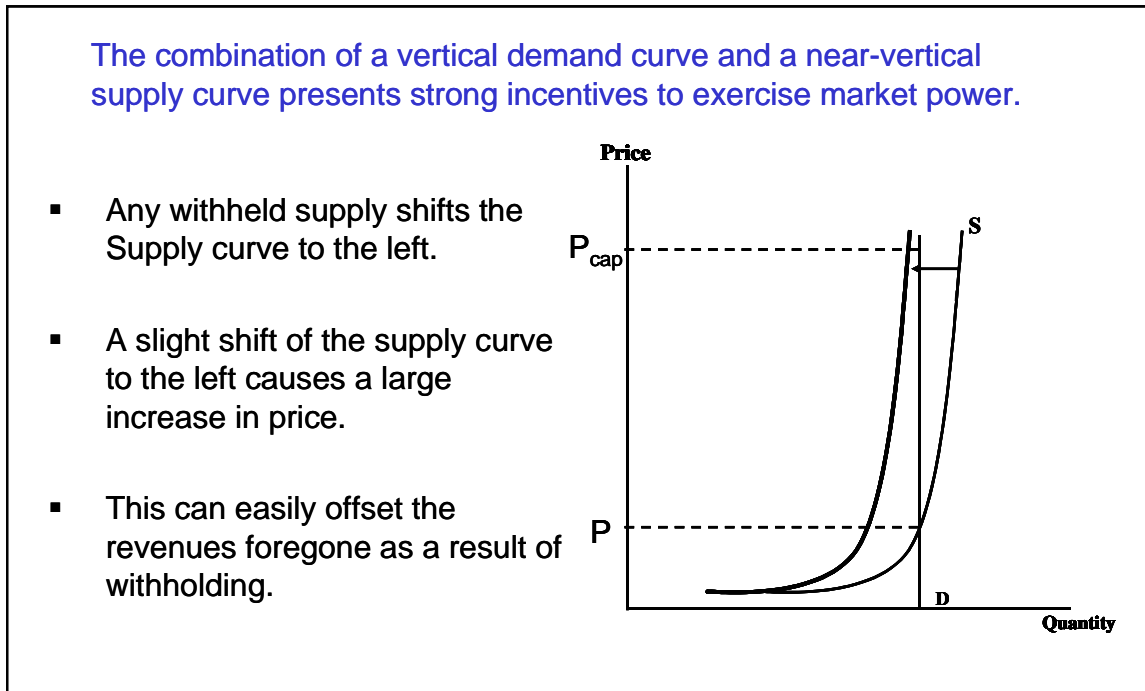
Figure 4
Price Volatility in Earlier ICAP Markets



Extreme price volatility caused by minor changes in supplies created high risks for buyers and investors, which tended to discourage contracting and investments in new construction. For example, while prices might be at price cap levels one month, the addition of a new plant might force prices down near zero, making recovery of the investors' fixed/capital costs uncertain.

The second problem created by the fixed, vertical demand curve was the incentive it created to exercise market power, depicted in Figure 5. Large suppliers could see that if they withheld only a small amount of their capacity from the monthly auctions, they could move the supply curve to the left just enough to force prices to leap from very low to very high levels. Capacity buyers would then be forced to pay very high capacity prices to the withholding supplier's remaining capacity, more than compensating the supplier for the capacity it withheld from the market.

Figure 5
Market Power Incentives in Early ICAP Markets



By 2002-2003, all of the ISO/RTOs recognized that the fixed, vertical demand curve for capacity was problematic; it encouraged both volatile prices and attempts to exercise market power to drive up capacity prices. Unstable prices were in turn discouraging investment, or at least raising the risks and costs of building new capacity. This common recognition led first New York ISO, and then New England ISO and PJM to rethink how they represented the demand for capacity.

In the past several years, PJM and other ISOs developed two basic remedies to solve the problems created by fixed, vertical demand curves:

- (1) *Change the demand curve.* Instead of using a fixed, vertical demand curve set at the required planning reserve margins (e.g., 15 percent), New York developed a downward sloping demand curve (see Figure 6) that represents a range of possible reserve margins centered around the 15 percent target. The effect of a downward-sloping demand curve is to reduce price volatility and to reduce the profits (incentives) capacity owners have to withhold capacity to exercise market power. New York has used downward-sloping demand curves for the last three years, which are explained further below. *PJM's RPM also uses a variation of the sloped demand curve approach.*
- (2) *Extend the auction to expand the supply curve.* One way to combat market power is through competition from new entry. To do this, an ISO can redefine the period in which capacity must be available. Instead of a monthly auction for capacity to be available the following month, the ISO can use annual auctions,

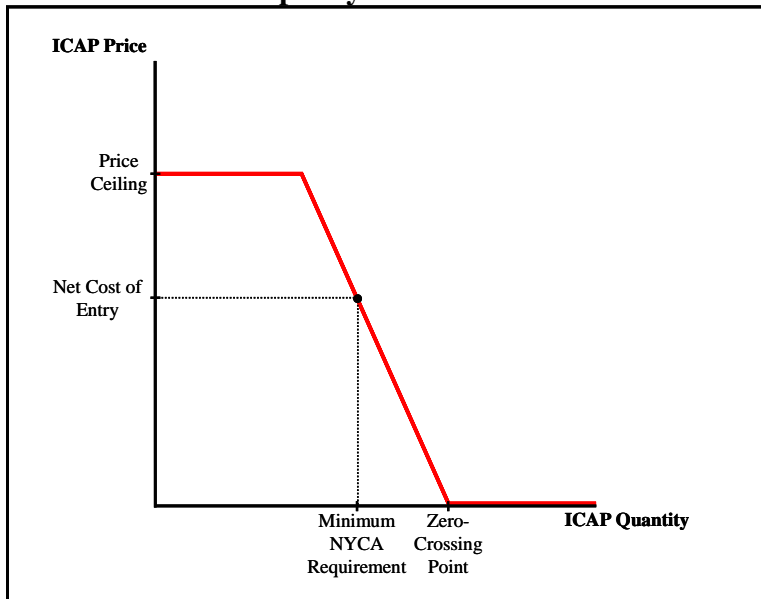
held 3-4 years in advance of when availability is required. Given this lead time, newly constructed capacity (and equivalent demand-side responses) can enter the market and counteract efforts by existing generators to exercise market power by withholding capacity. New England ISO adopted this forward auction approach in 2006; *PJM's RPM uses a similar forward auction approach.*

In sum, PJM's RPM uses the concepts from both solutions: it uses a variation of the downward-sloping demand curve pioneered by New York ISO, and it uses annual 3-year forward auctions like those eventually adopted by ISO-New England.

Development of Capacity Demand Curves and How They Limit Capacity Payments to Cost-of-Service Revenue Requirements

Figure 6 illustrates the concepts behind the original New York ISO demand curve.

Figure 6
New York ISO's Capacity Demand Curve



The New York ISO demand curve does not purport to represent buyers'/consumers' willingness to pay for capacity. Their willingness to pay is difficult to ascertain, because until recently most end use consumers were not metered in ways that could measure their hourly electric usage nor were there ways for them to know the changing spot prices for serving them so that they could adjust their usage depending on their willingness to be served at each price. Advances in metering technology are changing this over time, but for now, the ISO (or states) must administratively define the demand curve for reliability.

The default for determining the level of desired reliability has been to rely on long-established engineering practices that call for meeting a 1-day in 10-year loss of load expectation (LOLE) standard. In Figure 6 for New York, that standard is represented by

the “minimum New York Control Area requirement,” approximately a 15 percent planning reserve margin.

Given this reserve margin target, New York ISO worked with the New York Public Service Commission to construct a demand curve around a central point, as shown in Figure 6. This point represents the intersection of (1) the net cost of new entry – defined as the capital costs of a new gas-fired combustion turbine generator – and (2) the required 15 percent reserve target, which satisfies the Minimum New York Control Area Requirement. The “net” refers to the fact that the cost of entry is net of any profit margins the unit would be expected to earn from sales of energy and operating reserves in the NY ISO spot markets. In other words, the net cost of entry is the net amount of fixed/capital costs a combustion turbine would need to receive in the form of capacity payments to break even on a new investment, after accounting for expected profits from sales of energy and operating reserves.¹¹

Once this central point is established, the NY ISO needs two more points to define the curve. First, although some excess capacity has reliability value, there is some point at which there is so much surplus capacity (above 15 percent) that loads should be unwilling to make any further capacity payments. In Figure 6, this is the Zero Crossing Point, and it occurs wherever the ISO and State set it – in this case at about 20 percent reserve margins. Second, loads may be unwilling to pay above a price ceiling, no matter how short they are on capacity. In the figure, this Price Ceiling – a price cap on capacity payments -- was set at 1.5 times the Net Cost of Entry.

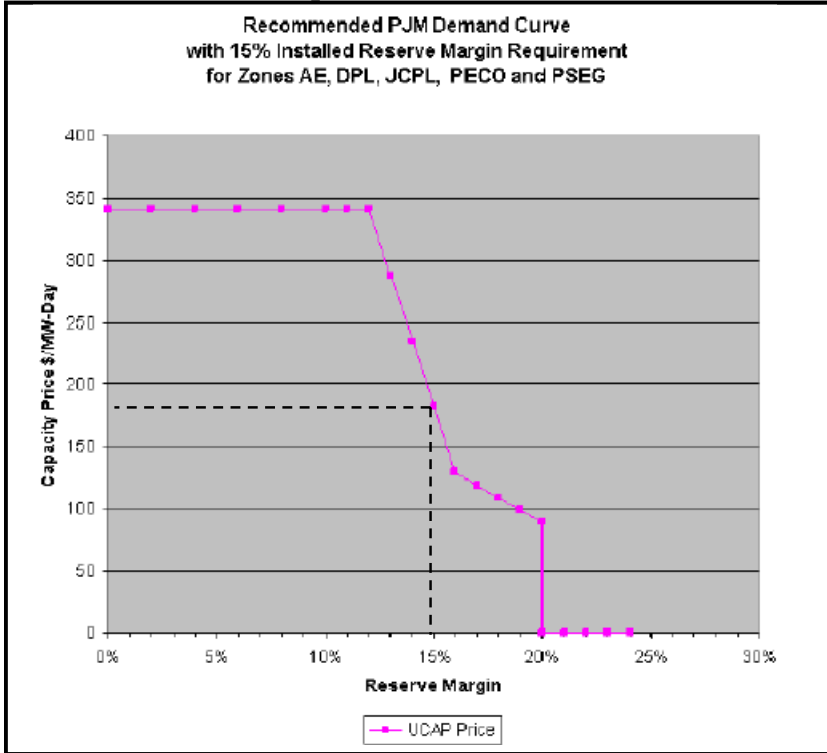
One way to understand the central point on the curve is that it represents the break even point for new investments when the total system just meets the reserve margin requirement. If there were more capacity available than the target 15 percent, then the demand curve would lead to capacity payments lower than the break even point, which would discourage further investments and might even encourage retirements of older generating plants that are expensive to maintain in service. Conversely, if there were less capacity available than the target 15 percent, then capacity payments would be higher than the break even point, which would tend to encourage new investments in new or refurbished units. The higher the surplus, the more likely owners are to retire the most expensive existing generating capacity to maintain, thus reducing the surplus; the higher the shortage, the more likely the higher prices will stimulate new plant investments, thus reducing the shortage. The sloping demand curve represents the range of prices at varying levels of capacity shortage or surplus.

¹¹ New York ISO has four types of operating reserve markets so generators can earn margins from both energy and operating reserve markets. In addition, New York integrates energy and operating reserve prices, such that shortages in operating reserves can cause energy prices to rise towards shortage-cost levels. In theory, this basic form of shortage-cost pricing should reduce the level of “missing money” and hence reduce the need for capacity payments compared to markets without these features.

PJM's Demand Curve

RPM's curve is patterned after the same concepts found in the New York approach, plus additional ideas taken from an earlier ISO-New England proposal. (New England eventually chose not to use its proposed downward sloping curve, but the refinements it developed appear in the PJM curve, as explained below.)

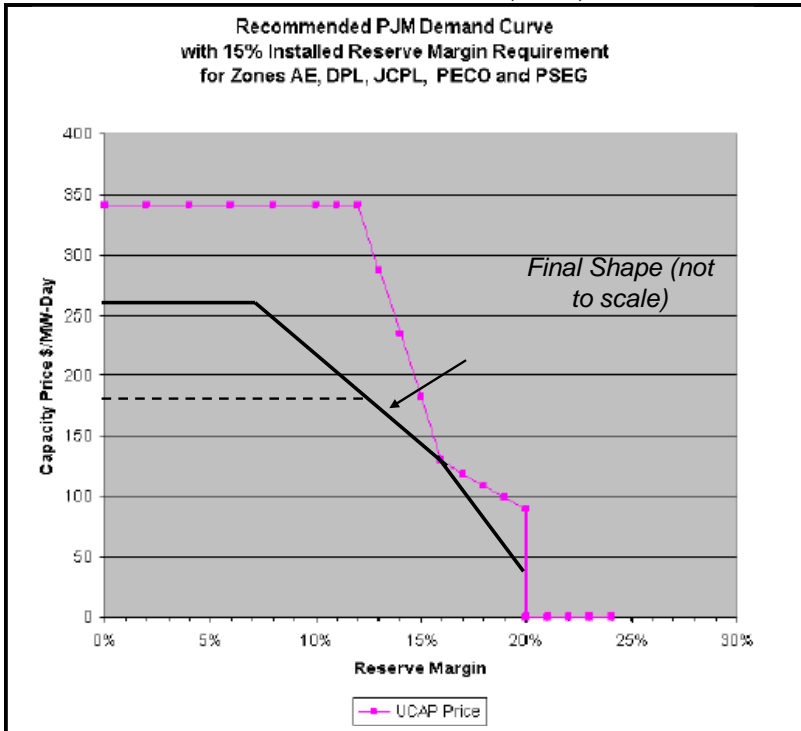
Figure 7
PJM Initial Demand Curve for
Variable Resource Requirement



In PJM's original proposal for a demand curve (Figure 7), the intersection of the break even point for new entry (shown by the horizontal dotted line) and the reserve margin target (about 15 percent, shown by the vertical dotted line) defines the central point around which the curve was originally constructed. As with NY ISO, there is a "zero crossing point" at which payments would be zero, and this occurs at 20 percent reserve margins. And there is a price ceiling or cap on capacity payments, which PJM originally proposed to be twice the net cost of new entry (Net CONE).

Figure 8 shows the final PJM demand curve that emerged from settlement discussions with other parties and approved by FERC.

Figure 8
PJM Final Demand Curve for RPM (2007)



Several aspects of the curve changed from the original proposal, and these changes tended to reduce the level of payments that will occur under RPM.

- The original “kink”¹² in the curve, which occurs when the amount of capacity is equal to about 16 percent reserve margins, has been turned up-side down. By lowering the curve in this manner, this change reduces payments under RPM compared to the original proposal.
- The price ceiling was lowered from 2 times Net CONE to 1.5 times Net CONE. This also lowers the curve and thus lowers payments during periods when the region is short of the desired level of reserves. (1.5 times Net CONE is also used as the price ceiling in New York)

¹² The idea for a “kink” came from the New England proposal. The rationale is that an ideal curve should have a steeper slope on the left side of the break even point and a shallower slope on the right side of that point. In New England the steeper slope on the left was designed to encourage investments when the amount of capacity is less than the desired reserve margins. The shallower slope on the right was designed to reduce the risks of slight over-investments, when capacity is close to the break even point. Inverting the kink could be argued to have the opposite effects, but without more experience, these effects are uncertain.

- Because the “kink” was inverted, the “break even” point for new investment probably occurs at some reserve margin less than 15 percent. This means that over time, the amount of investments will likely be somewhat below 15 percent, on average, because the amount of revenues available from RPM payments will be somewhat less than that needed for investors to break even on their investments if reserves are at 15 percent. Investors will likely build slightly less than they would have under the original proposal. However, this change is likely small and it is unclear whether consumers will notice any overall impact on capacity-related reliability.

Does the RPM Demand Curve Increase Total Costs for Ratepayers in the Long Run?

A frequent concern is that the RPM demand curve approach increases payments to generators above what they would be under a full cost-of-service regulatory regime. However, this is a misunderstanding of how the RPM demand curves are structured.

The cost of new entry (CONE) that PJM estimated represents the capital costs of installing a new combustion turbine, arguably one of the lowest cost sources of new generating capacity. Using independent experts to ascertain costs facing generation developers, PJM and its Market Monitor independently estimated the value of CONE in the same way and with the same kinds of information that state regulators might use to determine cost of service for new capacity. From a regulatory cost-of-service perspective, this approach means that the cost-of-service revenue requirements for the target level of reserves defines and limits the RPM payments.

Initial estimates of CONE may turn out to be too high or too low, just as regulators can under- or over-estimate a regulated utility’s generation cost of service when setting retail rates. If additional experience (as might be revealed by winning bids from successive RPM auctions) indicates that PJM’s estimated CONE differs from the actual CONE, the RPM rules provide a mechanism to readjust CONE to reflect this information. Given these adjustments over time, estimated CONE will tend towards actual CONE. In that way, RPM’s payments will tend to converge on actual cost-of-service, just as would occur with reevaluations of cost-of-service under a well functioning regulatory regime. However, RPM’s competitive features will tend to encourage efficiency improvements by plant owners and operators, so RPM is likely to lower cost of service compared to a strict regulatory regime.

As is true in New York mechanism, PJM’s RPM payments are reduced by the profit margins that resources receive from energy and operating reserve markets. Over time, if profits from energy and operating reserve markets go up, Net CONE is adjusted down and the demand curve adjusted downward, which means that payments under RPM will be lower in subsequent auctions. Conversely, if profit margins from other markets

decrease, Net CONE is adjusted up, raising the demand curve and increasing RPM payments in subsequent auctions.

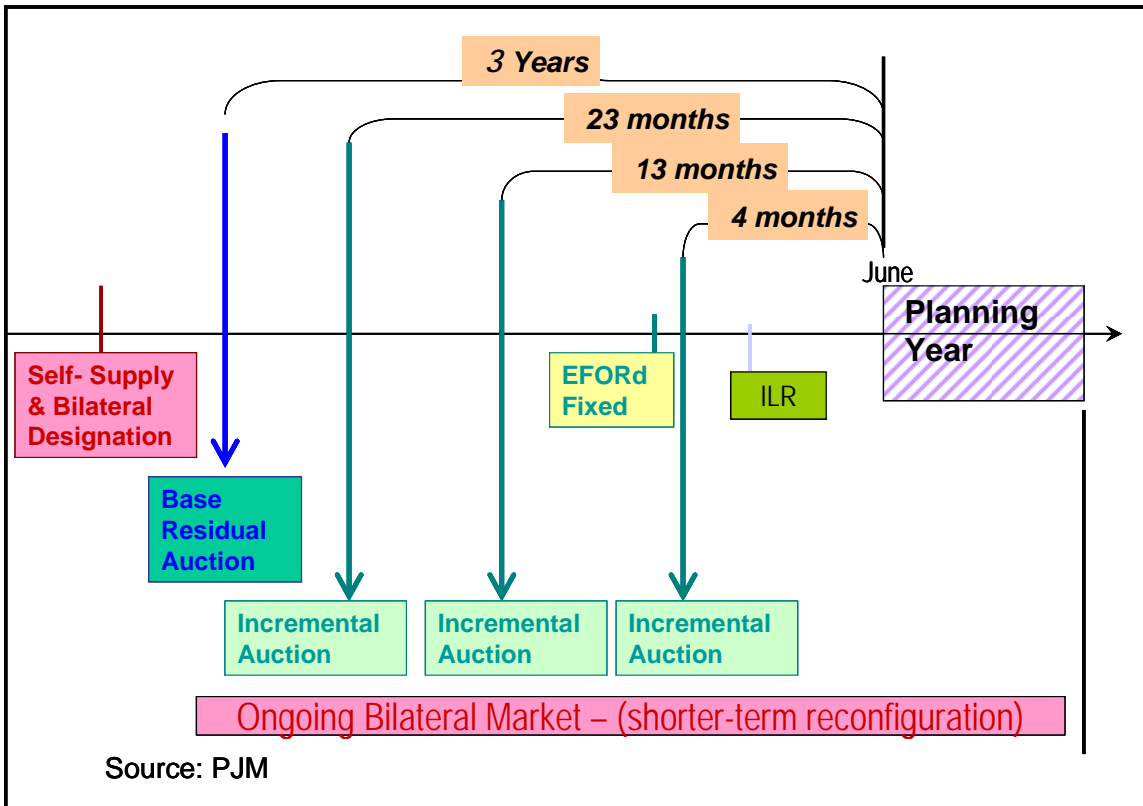
The competitive market aspects of RPM include the use of competitive auctions in which many capacity providers compete to supply the region's needed planning reserve margins; the competitive auctions ensure that the lowest cost providers are selected. In the long run, this part market/part cost-of-service approach means that the total payments made under the RPM mechanism will not be higher than what they would have been if the same level of resources were acquired under traditional cost-of-service regulation. In effect, and over time, cost of service sets a cap on RPM payments for whatever level of planning reserves the region selects.

Using Forward Auctions to Encourage New Investments

RPM's demand curves are constructed around a level of capacity payments that would be necessary to cover the cost of new entry, after accounting for estimated profit margins from sales of energy and operating reserves. When the level of capacity resources cleared in an RPM auction is less than the level of reserve margins required by the regional standards, RPM payments will be above the estimated Net CONE. Since Net CONE is the estimated break even point for new investment, the payments should, over time, provide an incentive for investors to add new resources.

However, the structure of RPM's forward auctions provides additional inducements for needed investments. RPM auctions are designed to acquire capacity resources from one to three years in the future, with the base auctions held three years in advance. These forward auctions are intended to provide sufficient time for new projects to be permitted and constructed, knowing they will receive RPM payments if/when they come on line as planned. The effect is to encourage new entry and to allow new entry to compete with existing capacity in the auctions to meet the regional and sub-regional reserve margin requirements. The basic approach of forward auctions is shown in Figure 9.

Figure 9
Timing of RPM Forward Auctions



Under this design, each year a base residual auction is held three years in advance of the planning year. Resource providers submit their offers in that auction, and the auction selects winning suppliers and determines the price for that auction, to be paid when the capacity resource becomes available in the planning year. If there are transmission limits between local deliverability areas (LDAs – see Figure 1), the auctions may determine different prices for each LDA, with prices higher in transmission-limited LDAs.

Conditions may change from year to year, so RPM provides several “incremental auctions” to account for these changes. For example, changes in expected retirements or changes in demand forecasts, as well as delays in permitting or construction of new plants, may increase the need for resources in the planning year. If that occurs, additional resources needed are acquired in the incremental auctions.

RPM also provides an opportunity for demand-response load reductions to earn RPM payments. RPM allows proposals for demand response to be submitted within the auction, but they also get a last chance to earn payments if offered in the months immediately before the planning year starts (ILR). This allows load reduction proposals with relatively short lead times to participate in meeting the region’s capacity needs.

RPM rules provide that selected resource providers will be paid the RPM payment determined in the annual auction for the planning year. In addition, RPM rules provide

for extended, multi-year payments in order to assure a steady payment stream for new resources for two additional years, to reduce investment risks. These payments, however, are not guaranteed and may be reduced or ended if a selected resource fails to meet the performance standards, such as by not being available during periods of operating reserve shortages or allowing its average availability to suffer.

Resources with lead times longer than three years for permitting/construction are not precluded, even though RPM uses a three-year forward auction construct. A three-year lead time may be enough for simple-cycle gas turbine additions, but probably not long enough for combined cycle configurations, let alone larger coal and nuclear facilities, to acquire permits and complete construction. Nevertheless, the economic underpinnings of RPM provide incentives for investments in plants with longer lead times *provided* investors have a reasonable expectation that RPM's basic structure will remain in effect and not undergo radical changes in design philosophy over an extended period.

RPM's rules allow adjustments to estimates of CONE and adjustments to estimates of the margins earned from energy and operating reserve markets. Investors can reasonably anticipate such changes and will factor the risks into their investment decisions. But the core of RPM is that RPM payments will, year after year, be based on the concept of paying Net CONE, so that the demand curve that determines payments will continue to be constructed around a central point that is logically related to investment revenue requirements. If this concept were abandoned or compromised, it is not clear how major investments could occur, since investors would have no assurance that the markets would support the level of investments needed to meet the region's reliability standards.

Regulatory certainty is thus critical to the success of RPM or any capacity payment mechanism. If investors believe the mechanism will remain in effect and function without radical redesign, the risks of making long-run investment decisions are reduced and market-driven investments are more likely to occur. If investors believe regulators are likely to intervene if capacity additions do not occur when regulators would prefer, then investment risks increase and market-based investments become much less likely.

Initial RPM auctions are thus likely to reflect some uncertainty and assessment of risks on the part of market investors, as investors gauge how well the mechanism is working. In the first two auctions, held for the planning years 2007-2008 and 2008-2009, auction prices tended to fall below the break even point for new investments. As long as prices remain below the break-even point, one should not expect new proposals for capacity addition. In the most recent auction for the planning year 2009-2010, prices in some LPAs are somewhat above the break even point.

Actual decisions to propose new capacity are of course affected by many factors, including siting/permitting difficulties, perceptions of state regulatory policies and conditions in capital and fuel markets. However, if this pattern of increasing RPM prices continues in the next year's auctions, one should expect to see the market responding with new proposals for capacity expansion. The results should provide a better indication of how well the RPM approach is working.

Summary of Features for PJM’s Reliability Pricing Model.

PJM’s RPM has many advance features, including these core components:

- (1) It defines forward capacity products that must be delivered up to three years in the future. This allows new capacity to enter the market and compete with existing capacity to meet the resource demand and set prices;
- (2) It uses annual bid-based auctions to acquire these forward capacity resources;
- (3) It acquires capacity resources on a sub-regional basis, reflecting the transmission limits that may prevent some distant resources from meeting local resource adequacy requirements; this feature may result in different RPM payments made to resources in different sub-regions of the PJM footprint;
- (4) It uses an administratively determined, downward sloping “demand curve” that defines the demand for capacity resources in each sub-regional market; the shape of the curve helps determine the price winning suppliers will receive;
- (5) RPM’s demand curve is structured around the cost of service for the least expensive capacity to build; the use of cost-of-service principles to define the demand curve has the effect of capping total market and RPM payment over time to levels no higher than would occur under cost-of-service regulation;
- (6) The RPM demand curve will be adjusted over time to reflect experience and bids in the annual auctions. This is a self-correcting mechanism that should tend to move investments towards the desired reserve level over time;
- (7) RPM auctions take account of investments in both generation and transmission; transmission expansions that increase transfer capability between LDAs are then reflected in the assumptions that define the amount of capacity that can be imported versus capacity that must be based inside an LDA; conversely, generation development in different LDAs influences PJMs transmission planning process and recommendations; RPM is thus part of a regional “integrated planning” process coordinated by PJM;
- (8) PJM applies market power mitigation measures to resources that wish to submit supply offers in the annual auctions; these measures prevent unjustified offers and penalize suppliers that unfairly withhold supply from the auctions;
- (9) RPM includes a set of administrative rules that, among other things, set the conditions capacity resource providers must meet to receive payments; these rules tend to encourage resources to be available when most needed, as when the PJM ISO has fewer operating resources than desired for reliable operations.

ⁱ John Chandley is a Principal at LECG, an economics consulting firm, and a member of a team of experts who advise on the design and operation of electricity markets and the role of Regional Transmission Organizations. This paper is part of a series of papers supported by PJM and designed to explain what RTOs do and how they function to support reliability and markets. The views expressed in this paper, and any errors, are solely the responsibility of the author.