

The Demand for Operating Reserves: Key to Price Spikes and Investment

Steven Stoft

Abstract—Under regulation, operating-reserve policy and investment policy are completely separate. In a market, they are tightly linked through expectations. Currently, regulators and engineers intervene in markets to determine how much will be paid when operating reserves are in short supply. These prices determine the revenue stream that pays the capital costs of new peakers and pays an equal amount toward the capital costs of all other generators. In this way, operating-reserve and price-cap policies determine investment in generation and the equilibrium level of installed capacity. Typically, FERC determines a price limit, and engineers, by setting an operating-reserve requirement, determine the expected duration of price spikes. Currently, these policies are set without coordination and without analysis of the long-run consequences.

Index Terms—Price spikes, investment, installed capacity, operating-reserve requirements, long-run market equilibrium.

I. INTRODUCTION

CURRENT power markets suffer from extreme price spikes and poorly planned investment in generation. The connection is obvious: price spikes induce investment, but they are erratic. California had virtually none for two years and then had nothing but price spikes for almost an entire year. While this case is extreme, even PJM's cumulative annual price spike is quite erratic.

The discussion of the investment problem typically follows four steps:

1. We should “let the market work.”
2. But we need “damage-control” price caps.
3. But that will cause inadequate investment leading to reliability problems.
4. So we need an installed-capacity requirement.

Step 1 is currently irresponsible. Step 2 is so vague it provides little guidance. Step 3 is simply wrong. Step 4 assumes the truth of step 3. This paper uses the theory of competitive markets to model the investment process and demonstrate the following.

1. Currently “the market” generates no reliability information and cannot approximate optimal investment without external guidance.
2. Price caps are useful for reducing volatility and market power.
3. They may not, and need not, cause under-investment. That depends on the operating-reserve requirement.

4. Policy makers could choose from an array of optimal-investment policy options if engineers calculated those options from known market data.

The root of the confusion is a failure to comprehend the full impact of demand side flaws. The lack of demand responsiveness to price is known to exacerbate market power dramatically. Its effect, in the absence of market power, on investment-inducing price spikes has not been understood.

If consumers were faced with a real-time prices, one could almost imagine that their behavior in the face of extreme prices would signal the consumer's value and something about their desire for reliability. But most consumers receive no credit for curtailing their use when the spot price is high. Thus, the market learns nothing from high spot prices about consumer preferences for reliability. Most consumers never consider the question of what reliability is worth to them. Even if a market could read minds, it could not determine optimal investment. The required information simply does not exist. Thus the high prices observed in present-day power markets do not reflect the desire of consumers for reliability, but instead reflect short-run regulatory and engineering policy.

Because the market cannot do the job, engineers and regulators have stepped into the breach. They control investment—though never in a calculated manner—by setting price caps, operating-reserve (OpRes) requirements, installed capacity (ICap) requirements, and ICap penalties. This paper analyzes effects of price-caps and OpRes requirements.

This paper assumes an energy market exists and, except when noted, assumes suppliers do not exercise market power. It assumes engineers will set minimum requirements for reliability and regulators will cap prices, and then investigates how these policies should be coordinated to produce efficient un-regulated investment. It does not analyze ICap requirements though it suggests they can be useful in present markets, and presents a framework that can easily include them [1 at 180].

II. PRICE CAPS AND OPERATING RESERVE REQUIREMENTS

Regulators, especially the Federal Energy Regulatory Commission (FERC), control price caps, while engineers control OpRes requirements. While price caps are known to modify the demand function and profoundly influence price spikes and investment, the similar impact of OpRes requirements is generally overlooked. OpRes requirements are needed and designed to provide short-run system security—one component of reliability. That function is not in question.

In a regulated environment, that is their sole function, but in a market, they affect prices in the energy market and, through these, affect the long-run equilibrium level of installed capacity. While the OpRes requirement is implemented solely for its short-run reliability benefit, it has an equally important long-run investment impact, which is often overlooked, especially in the United States. To understand this effect, two aspects of OpRes requirements must be examined: (1) enforcement of the requirement, and (2) arbitrage between the reserve markets and the energy market.

In a regulated system, OpRes requirements are specified simply by a level. At a particular time, the spinning reserve requirement may be 2500 MW. If the requirement can be met it is met; if it cannot be, it is not. There is no question of how hard the system operator should try to meet the requirement.

A. A Price Limit for Operating Reserves

In a market, the system operator must purchase OpRes, which introduces the question of price. Were prices always “reasonable,” say under \$200/MWh, the old approach could be maintained. Just meet the requirement if possible.

For the sake of concreteness, assume that the total OpRes requirement, counting regulation, spinning reserves and non-spinning (replacement) reserves, is 10% of load. What should be done if, due to a shortage of replacement reserves, the amount of OpRes available at a reasonable price drops to 9% and the additional 1% of replacement reserves could only be purchased at a price of \$10,000/MWh. The CA-ISO faced this price in July, 1998 and made the purchase.

Such decisions are usually justified by the regulatory approach of “meet the requirement if possible.” Can this approach make sense in a market? OpRes requirements, though based on sound principles are rules of thumb that vary from one control area to another. Is it possible that each MW of OpRes is worth \$10,000 up to the requirement, but the next MW is worth nothing? Is it possible that there is no limit on what the system operator should pay to meet the last MW of the requirement?

In more recent years, the CA-ISO has shown much greater flexibility in violating its operating reserve requirement and has never again paid \$10,000/MWh even with reserves running very low. Though much could be written concerning the correct pricing of OpRes from a short-run reliability perspective, the purpose of this paper is only to analyze the long-run implications of such pricing. For that purpose, assume the OpRes requirement is a function that specifies how much the system operator should offer to pay at every level of OpRes. It is sufficient for the present analysis to imagine that nothing will be paid for reserves above the OpRes Requirement (OR^R) and that the price cap (P_{cap}) will be offered whenever OpRes falls below OR^R . Note that the price cap is simply a limit on how much the system operator will offer to pay.

B. Arbitrage between Energy and Reserve Markets

Many generators have a choice—they could sell less energy and more reserves or vice versa. Choosing between these markets is a form of arbitrage. Choosing to sell energy instead of reserves saves roughly the marginal cost, MC , of energy.

Arbitrage between the two markets will keep the price of energy (P) approximately equal to the price of reserves (P^R) plus the marginal cost of producing energy.

$$P \approx P^R + MC \quad (1)$$

Consequently, setting the OpRes price can have a much greater indirect effect on profits through the energy price than it has directly through the price of reserves. A second arbitrage relationship magnifies this impact.

The day-ahead energy market anticipates the real-time (spot) market. If traders believe the spot price will be \$30/MWh, they will not pay \$40/MWh in the day-ahead market but will wait for real time to buy power. This brings a high day-ahead price down toward the spot price. Similarly, they would snap up \$25 power in the day-ahead market, and this brings the day-ahead price up towards the \$30 spot price. Thus, arbitrage equates the day-ahead price with the expected spot price. This standard economic prediction has been checked in all of the ISOs and found to hold quite accurately.

The same logic applies to all forward markets. Aside from minor discrepancies due to risk, the forward price is the expected future spot price. Consequently, all forward energy prices are approximately as great as the expected price of reserves plus the variable cost of production.

To a reasonable approximation, the short-run profit of a generator that supplies energy is the price of energy minus the marginal cost of production. Because the marginal cost in (1) is system marginal cost, which may be higher than the cost of base-load generators that have backed off to supply reserves, short-run profit (SR_π) is greater than or equal to the price of operating reserves.

$$\begin{aligned} SR_\pi &\geq P^R \\ SR_\pi &\approx P^R \quad \text{for a peaker} \end{aligned} \quad (2)$$

For peakers, whose marginal cost (or average variable cost) is approximately system marginal cost, there is near equality between short-run profits and the price of reserves.

Various effects contribute to the discrepancy, but these typically amount to a few dollars per MWh and the effects considered in this paper typically depend on short-run profits that are more than \$100/MWh. So, while more accuracy may be desirable at a later stage, the simple approximation in (2) will be taken as an equality in the following discussion.

For power sold in forward markets, short-run profits for peakers also equal the expected price of reserves. The average price of reserves can thus be used to calculate average short-run profits (for peakers) from all markets, and these drive the investment decision.

III. A SIMPLE LONG-RUN MODEL

The expectations of future short-run profits from high OpRes and energy prices control investment. This is due to the arbitrage relationships between short- and long-run markets just described. Thus expectations of high spot prices lead to high forward prices and, no matter which prices the investor considers, to high investment. But investment affects capacity

and capacity affects prices and thus profits. The short-run profit function connects capacity to profits.

A. The Short-Run Profit Function

The break-even condition for generation investment is that expected short-run profits equal fixed costs. Higher profits will attract new investment and lower profits will put a stop to investment. Investment increases the level of ICap, and lack of investment allows it to dwindle relative to load.

The short-run profit function links ICap to short-run profits, which completes the negative feedback loop that determines what economists call the long-run equilibrium. This feedback can be diagrammed as follows:

$$\text{ICap} \uparrow \rightarrow \text{SR profit} \downarrow \rightarrow \text{Investment} \downarrow \rightarrow \text{ICap} \downarrow$$

When installed generation capacity is high, the supply of energy is increased and this depresses the average price. A reduction in the price received by generators necessarily reduces their profits. To the extent these conditions are expected to continue, lower profits will be anticipated, and this will discourage the construction of new plants whether by new or existing players. But with investment down, plant retirements and load growth will eventually reduce the level of installed capacity relative to peak load. This will cause a shortage, raise prices and induce more investment.

The short-run profit function, $\text{SR}_\pi(K)$, describes how expected short-run profits depend on ICap, K . The expectation is taken over generation outages, weather, and other sources of random load fluctuations. It also depends on two policy parameters, P_{cap} and OR^R . Although each generation technology has its own profit function, because of (2), it will be most convenient to focus on the short-run profit function of peaker technology.

B. The Model's Equations

A simple example illustrates the basic principles. The model requires demand and supply sides, the regulatory rules, and the market's equilibrium condition. Its first equation is the load duration curve modified to take account of generation outages. These are counted as load to help explain why the target value of ICap is roughly 118% of peak load. Load plus generation out of service will be called "augmented load" or "load" and denoted by L_g . "Normal load" will refer to load exclusive of generation out of service. The first model equation describes the load duration curve shown in Fig. 1.

$$L_g = L_g(D) = L_{\text{max}} - \alpha D \quad (3)$$

Duration, D , is measured as a fraction of a year, thus the duration of baseload operation is almost one. Since load is assumed to be completely inelastic with respect to price, this completes the description of the demand side. Note that (3) only describes the load-duration curve near its peak. This is sufficient for the computation of peaker profits.

On the supply side, assume the only generation technology has a marginal cost, MC , and a fixed cost, FC , but no startup or no-load cost. Generator profits depend on prices as well as cost, so the next two equations describe price regulation.

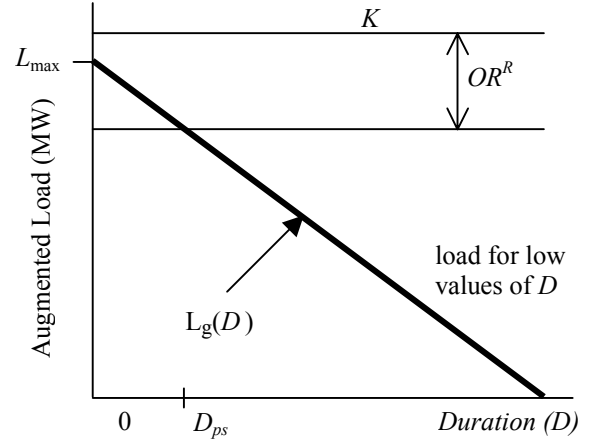


Fig. 1. Load, capacity, and operating reserves in the inelastic model.

When $L_g + OR^R < K$, there is no shortage of operating reserves, and P^R is zero. When $L_g + OR^R > K$, then P^R is set to the price cap, P_{cap} , and there is a "price spike."¹ The annual accumulation of price spikes, which can be depicted with a price duration curve, will be called the aggregate price spike. Its duration is called D_{ps} , and can be found from:

$$L_g(D_{ps}) + OR^R = K \quad (4)$$

From (2), SR_π equals P_{cap} during a price spike, so expected short-run profits are given by P_{cap} times the duration that the market price is at the cap. Consequently the short-run profit function is:

$$\text{SR}_\pi(K) = D_{ps}(K) \times P_{\text{cap}}, \quad (5)$$

This completes the description of the supply side.

The remaining equation is the market's long-run equilibrium condition, which states that expected short-run profits must equal fixed costs. Because SR_π is the average \$/MWh of installed capacity, FC must be amortized and divided by the number of hours per year to convert it to \$/MWh. The equilibrium condition for peakers (which are the only generators that plan an explicit role in this model) is then:

$$\text{SR}_\pi(K^E) = FC \quad (6)$$

This equation determines the long-run equilibrium value of ICap, K^E , and completes the specification of the model.

C. Solving the Model

The first step toward solving the model is to find how D_{ps} depends on K , by substituting (3) into (4) and solving for D_{ps} , which is then acknowledged to be a function of K . This gives:

$$D_{ps}(K) = (L_{\text{max}} + OR - K) / \alpha \quad (7)$$

Next, (7) is used with (5) to find the short-run profit function:

$$\text{SR}_\pi(K) = (L_{\text{max}} + OR - K) P_{\text{cap}} / \alpha \quad (8)$$

¹ For algebraic convenience, price-caps are assumed to limit the price offered for reserves rather than for energy. Due to arbitrage, the two approaches provide an equivalent set of policy options.

The final step is to find the level of K that produces expected short-run profits that just cover fixed costs by substituting (8) into (6) and solving for K^E :

$$K^E = L_{\max} + OR - \alpha FC / P_{\text{cap}} \quad (9)$$

TABLE I
DEFINITIONS OF VARIABLES AND THEIR UNITS

Demand side of market

L_g	MW	Load + generation out of service	
D	—	Duration of load	
L_{\max}	MW	Maximum L_g	8000
α	MW	$L_g(D) = L_{\max} - \alpha D$	50,000

Supply side of market

MC	\$/MWh	Marginal cost of generation	20
FC	\$/MWh	Fixed cost of generation	10

Regulatory rules

P_{cap}	\$/MWh	Price of reserves when $L_g > K$	1000
OR^R	MW	OpRes requirement	800

Market outcome (short run)

P	\$/MWh	Price of energy	
P^R	\$/MWh	Price of reserves	
OR	MW	Operating reserves (OpRes)	

Market outcome (long run)

D_{ps}	—	Duration of aggregate price spike	
SR_{π}	\$/MWh	Expected short run profits	
K	MW	Installed generation capacity (ICap)	
K^E, K^*	MW	Equilibrium and desired K	

Note: \$1/kW/year = \$(8765/1000)/MWh. FC and SR_{π} are measured in \$/MWh of capacity, not energy.

The duration of the aggregate price spike is found by substituting this value of K^E back into (7) to find:

$$D_{ps}(K^E) = FC / P_{\text{cap}} \quad (10)$$

Having solved for the economic outcome—installed capacity, given the two policy parameters P_{cap} and OR^R , it is useful to turn this result around. Either policy variable can be found from the desired level of installed capacity, K^* , and the other policy variable. For example, (9) can be solved for OR^R . This gives the operating reserve requirement as a function of P_{cap} and K^* .

$$OR^R = K^* - L_{\max} + \alpha FC / P_{\text{cap}} \quad (11)$$

Returning to the computation of long-run equilibrium ICap, K^E , consider the values of the market-structure and regulatory parameters shown at the right of Table 1. Substituting these into (9) and (10) yields $K^E = 8300$ MW, and an equilibrium aggregate price spike duration of 1% or about 88 hours per year. This equilibrium is completely reliable because load plus

generation out of service reaches a maximum of only 8000 MW compared with an installed capacity of 8300 MW.

IV. THE PRICE-CAP FALLACY AND THE POLICY TRADE-OFF

Too low a price cap, or perhaps any price cap, will discourage investment and lead to an unreliable system—this is standard “wisdom.” But, what is “too low?” FERC just raised the Western price cap from \$98.37 to \$250/MWh to ensure enough investment. However, some Australian’s argue that optimal investment requires a price cap of \$10,000AU, their current value. The NYISO once had a FERC-approved price cap of \$10,000 for this purpose. A previous FERC Chairman argued, in the midst of California’s meltdown, for no price caps—he thought even the highest price cap was too low. FERC’s current chairman has referred to price-caps as “something from the kitchen slop bucket.”² The popular view is that any price cap will damage investment, and this requires a tradeoff. Thus, FERC is now approving “damage-control” price caps. Their theory of such caps is still murky to non-existent.

A. The Price-Cap Fallacy

In the present example, a price cap of \$1000 induces an equilibrium installed-capacity level of 8300 MW, 300 MW greater than any combination of load and outages ever experienced by the market. Within this model, the excess capacity is wasted. This price cap (along with OR^R) has induced too much investment no matter what the value of value of lost load (VOLL). A slightly lower price cap would save money without reducing reliability.

Fallacy: Any price cap below VOLL induces too little investment for system reliability.

Result 1: A price cap below VOLL can be too high.

Although a price cap of \$1000/MWh is too high, perhaps a price cap of \$250/MWh is too low. That depends on the OpRes requirement. Substituting $P_{\text{cap}} = \$250/\text{MWh}$ and $K^E = 8300$ MW into (11) gives an OpRes requirement of 2,300 MW. With this OR^R , a price cap of \$250 would induce the same ICap, which means there is still too much investment.

Result 2: Almost any price-cap level can induce too much investment.

Result 2a: Almost any price-cap level can induce too little investment.

Consider a real market instead of the model. While $P^R = \$250/\text{MWh}$, peakers are earning short-run profits well over \$150/MWh, i.e. they are covering their average fixed costs more than twenty times over. Say the target ICap is 20% above peak load. Setting OR^R to 80% above peak load will guarantee that a system with the target ICap is very often short of reserves and peakers are often earning over \$150/MWh of profit. This is far more than needed to induce the target ICap

² Reuters, March 5: Wood said that FERC's California order sent "bad signals" to industry players. But he offered a colorful defense of FERC's response: "When there's a fire in the kitchen you don't look at what's in the slop bucket. You just throw it." Chairman Wood has the most enlightened

level. If too much investment can be induced with an extreme OR^R , then the right amount can be induced with a more modest OR^R .

B. The Policy Trade-off

The optimal ICap can be achieved with any of a continuum of different policy options ranging from extremely high price caps and low OpRes requirements to very low price caps and high OpRes requirements. Short-run reliability considerations may rule out some policy options, but if so, it is the high price spike options with their minimal OpRes requirements that will be ruled out. Engineers may wish to rule-out a high OpRes requirement because it is inefficient to have a coal plant provide either spin or 30-minute non-spinning reserve if that is not needed for short-run reliability. But if an extra-large OpRes requirement were desired as an investment incentive, the increased requirement could be added in the form of a 24-hour non-spinning reserve requirement. This would leave the standard OpRes requirements unchanged, and cause no short-run inefficiency.

As shown by (10), the choice is between high, short-duration spikes and low, long-duration spikes. As long as (with optimal ICap) the product of height and duration equals fixed costs, the policy will induce optimal investment. Consequently, only the secondary effects of these policy choices can be used to choose between them.

The most obvious effect is volatility. Ten-thousand-dollar price spikes that occur once every three years and cause a near bankruptcy or two will grab more headlines than many \$300 price spikes that occur regularly throughout the summer. But volatility affects more than public opinion. A second effect is an increase in market power caused by high price spikes.

1) The Cost of Risk Imposed by High Price Spikes

High price spikes not only cause more day-to-day fluctuation in profits, they also cause more year-to-year fluctuation. A cool year may have none; a hot year may yield several years' profits in a few days. Such year-to-year volatility represents a real business risk because it makes the long-run average rate of return difficult to predict. Investors always demand a risk-premium on risky investments, so the cost of funds in a high-price-spike market will be higher. This is a real cost and may be significant. Currently, U.S. power markets are viewed as very risky, a fact that must be attributed largely to the effects of high price spikes and the related long periods of insufficient profit that bring SR_π down to the right long-run average. This level of perceived risk has a significant cost, especially when compared with the extremely low risk level that was achieved under regulation.

Some economists have argued that long-run contracts can overcome risk. Once a generator has sold all its power for the next ten years with a properly indexed contract, it has eliminated most of its market risk. Unfortunately, since California's Governor stopped purchasing emergency power with ten-year contracts, this market is once again thin to non-existent.

Even if most load-serving entities did attempt to purchase all of their power with ten-year contracts, multi-year swings in the spot price make it difficult to price such contracts accurately. This will make the long-run price risky and will produce boom-bust investment fluctuations. Both are costly.³

2) Increased Market Power Due to High Price Spikes

The second effect is less obvious. The ability to push prices to \$10,000 instead of \$300 certainly increases the exercise of market power when the market is tight. But, the market is genuinely tight only when a competitive market would be short, or nearly short, of operating reserves, and this happens much less frequently in the high-spike world with its low OpRes requirement. Preliminary investigations of these market power effects [1 at 171] indicate the long-run average cost of market power is higher in a high-price spike market.

V. WHY CAN'T THE MARKET DO ITS JOB

Many engineers consider it almost axiomatic that a market cannot induce the appropriate level of investment. But others, having converted to a market philosophy, believe that unfettered markets can solve nearly any problem. It is this blind faith that has so far prevented an understanding of the policy trade-off presented in this paper. This section uses standard competitive theory to explain why the flaws in current power markets prevent them from solving the investment problem without external guidance.

Currently, U.S. markets have price caps, which determine how high prices rise, and OpRes requirements that determine the duration of price increases.⁴ These regulatory and engineering rules thereby determine short-run profits and investment. Many claim that an unfettered market and the standard competitive mechanism would induce optimal investment, or at least come closer to that goal than engineers and regulators. Isn't this what economics teaches?

The answer is "Yes" if the demand side is working well enough, and "No" if it is not. Defining "well enough" proves difficult because a small quantitative shift in demand elasticity can eliminate the market's long-run equilibrium, and even before this extreme failure, it will make the equilibrium less efficient than a properly regulated one. First, consider the case of an ideal power market with a fully functioning demand side. Assume load is elastic and can be reduced to a low level by a high enough price.

Assume the demand for OpRes is expressed as a totally inelastic demand—an offer to pay any price. An ideal market buffers this irrational demand by raising the price of power until demand shrinks enough that the system operator's demand for reserves can be met. With sufficient demand elasticity, this will hold price below VOLL at all times. A model will best explain the mechanism of the ideal market.

³ Moreover, most forward markets depend on speculators (power marketers) for much of their liquidity. Speculator profit from beneficial arbitrage, but also require a risk premium. Extreme volatility will both raise this premium and disrupt the business of trading power.

⁴ A popular theory holds that price caps provide a target to shoot at for those exercising market power. That is not the effect considered here. Except for the paragraph on market power, this paper concerns only high prices caused by genuine supply shortages.

price cap policy to date, but colorful price-cap bashing may offer protection from market fundamentalists.

A. Equilibrium when the Demand Side Works

The “elastic model,” with a working demand-side, differs from the previous, inelastic, model in only two respects. First, demand is price elastic, and second the price cap is effectively infinite; in other words, the demand for OpRes is completely inelastic. Demand elasticity of load requires a modification of (3), and the infinite price cap requires a modification of (4). Equation (5) is modified only because the aggregate price spike is triangular in the elastic model, rather than rectangular. Equation (6) remains unchanged. The new versions of (3)—(5) are as follows:

$$L_g(D, P^R) = {}_E L_{\max} - \alpha D - \varepsilon P^R \quad (12)$$

$$L_{\max} = L_g(0, P_{\max}^R) = L_g(D_{ps}, 0) = K - OR^R \quad (13)$$

$$SR_{\pi}(K) = D_{ps}(K) \times P_{\max}^R(K) / 2 \quad (14)$$

Equation (12) includes an elasticity parameter, ε , which does not equal elasticity but which increases with it. The new parameter, ${}_E L_{\max}$, is the maximum possible value (accounting for elasticity) of augmented load, L_g , which corresponds to $D = 0$, and $P^R = 0$. This value of load never occurs because, when $D = 0$, P^R is at its maximum value, P_{\max}^R .

Equation (13) differs from (4) because load is constant from $D = 0$ to $D = D_{ps}$. The flatness of the load duration curve during the price spike (see Fig. 2.) is due to the inelasticity of the demand for OpRes and the elasticity of normal load. Whenever OpRes is threatened, price increases to the point where normal load backs off just enough to allow the system operator’s demand for OpRes to be fully satisfied.

Equation (14) computes the area of the aggregate price spike, which is triangular because the demand function is linear in both price and duration.

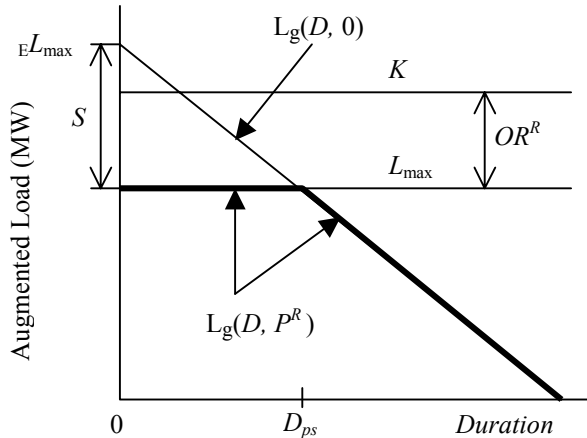


Fig. 2. Load, capacity, and operating reserves in the elastic model.

In this model it may be helpful to think of duration as if it were time and load as if it progressed smoothly from its peak value at time 0 to its minimum value at time 1, which indicates the end of the year. This causes no confusion because the model contains no startup costs or ramp-rate limits.

To solve this model it is convenient to define the capacity shortfall, S , to be the gap between the maximum possible

demand, ${}_E L_{\max}$, and the greatest load that can be served without running short of reserves, $K - OR^R$ (see Fig. 2.). Equation (13) indicates that the maximum actual load, L_{\max} , will be no greater. So, the shortfall is

$$S = {}_E L_{\max} - L_{\max}(K) \quad (15)$$

Next, evaluate (12) at two points, $D = 0$ and $D = D_{ps}$. In both cases, (13) indicates load is L_{\max} , so (12) reduces to

$$S(K) = \alpha D_{ps} + \varepsilon P_{\max}^R \quad (16)$$

When $D = D_{ps}$, $P_{\max}^R = 0$, so evaluation at the two points gives:

$$S(K) = \varepsilon P_{\max}^R \quad (17)$$

$$S(K) = \alpha D_{ps} \quad (18)$$

Using (17) and (18) to substitute for D_{ps} and P_{\max}^R in (14) gives the short-run profit function:

$$SR_{\pi}(K) = S(K)^2 / 2\alpha\varepsilon \quad (19)$$

This equation says that short-run profit increases when K decreases because the resulting increase in $S(K)$ increases the height and duration of the aggregate price spike. Using the equilibrium condition (6) gives the equilibrium value of the shortfall S :

$$S^E = \sqrt{2 \cdot \alpha \cdot \varepsilon \cdot FC} \quad (20)$$

Solving for K in (13) and substituting ${}_E L_{\max} - S$ for L_{\max} (from 15) gives the equilibrium value of K :

$$K^E = {}_E L_{\max} + OR^R - \sqrt{2 \cdot \alpha \cdot \varepsilon \cdot FC} \quad (21)$$

The equilibrium values of P_{\max}^R and D_p can be found by substituting S^E into (17) and (18).

If this example is evaluated using the same parameters as were used in the previous example and $\varepsilon = 1$ (meaning demand is reduced by 1 MW per \$1/MWh increase in price), the equilibrium value of ICap is found to be 7800 MW. This is less than the maximum possible demand, but the system is hyper-reliable; it is never short of operating reserves. The maximum price is, by coincidence again \$1000/MWh and the duration of the aggregate price spike is 2% instead of 1% because the average price of OpRes during the triangular price spike is only \$500/MWh.

If demand were more elastic, the load-duration curve were less sharply peaked, or the demand for operating reserves were more elastic, the price spike would be lower.

B. Pushing the Ideal Case to Its Limit

In the elastic model, the market clears without a price cap, and thereby removes the need for one of the two policy controls. However, the OpRes requirement is still in effect and still has a large influence on installed capacity. What has gone wrong? We have fixed the market and removed the price cap, but investment is still controlled by regulators—most likely of the engineering variety. What can be done about this?

Notice that in Figure 2, price elasticity always holds augmented demand, L_g , to a level low enough that the operating reserve requirement is fully met. This makes the

ideal power market hyper reliable relative to a normal regulated power system or even a power market that relies on paying VOLL when there is load shedding as called for by the theory behind Australia's power market. In both of these cases OpRes is short of the required value periodically and vanishes entirely about "one day in ten years."

Obviously the extra reliability of this ideal market is costly and not needed. This problem can be reduced by reducing the OpRes requirement, OR^R , which reduces equilibrium ICap and thus saves money. Within this model the reduction in ICap does not decrease reliability because augmented load never exceeds installed capacity (normal load never exceeds available capacity).

As OR^R is reduced, so is the role of engineering in determining investment and the long-run equilibrium level of installed capacity. If the ideal model is accepted literally, OR^R can be reduced to zero leaving price and demand elasticity to clear the market. This will keep demand from exceeding supply and provide reliability.

If demand elasticity fails to perform these functions perfectly, a small OpRes requirement may be needed to ensure reliability. If this is the case, the requirement will necessarily be violated at times, otherwise it would not have been needed. This necessitates setting a price to be paid when the violation occurs. The profits due to regulation will be determined by the frequency of these violations and the regulated price paid when they occur. If these profits constitute a small fraction of the fixed costs of peakers, then the engineering/regulatory policy can be said to play a small role in the determination of investment and reliability. The market is the main determinant.

Power markets in the U.S. presently cover the fixed costs of peakers (to the extent they are covered) almost entirely by revenues governed by engineering/regulatory mechanisms. If it were not for such mechanisms, profits would fall, investment would nearly cease and reliability would soon be drastically reduced. Present markets are far from the ideal just described.

In the future, demand elasticity may increase greatly and the market may, to some extent, approach the ideal just described. Three phenomena are likely to prevent the market from reaching this ideal. First, some operating reserve, "regulation," is needed to control frequency, which is a public good. Market's generally fail to adequately supply such goods. Second, some operating reserves may be needed to prevent cascading system failures. Again this is a public good, but the extent to which this problem could be prevented more economically (in a nearly-ideal market) with controlled load shedding has not been determined. The answer may depend on the extent to which normal operating reserves can be replaced with demand elasticity in such a market.

The third difficulty with approaching the ideal market may be an inherent lack of sufficient high-speed demand elasticity to handle normal contingencies such as sudden generation outages and line outages. While these could be handled with controlled load-shedding, that is probably not economical. In any case, it would require a regulated price since, by definition the supply and demand curves have failed to intersect (otherwise demand elasticity would have solved the problem).

Present markets are far from the ideal imagined by economists who advocate a complete reliance on the market. While this may seem obvious to many readers, one may be sure that the notion that the market should be left to solve these problems on its own will continue to thrive. Consequently it is useful to understand exactly how the market would fail, were it left unregulated.

C. Market Failure when Demand Elasticity is Too Low

As lower elasticities of demand are considered, the maximum price in the ideal model, P_{\max}^R , increases without limit, but while elasticity is finite, the model continues to have an equilibrium. However, the equilibrium can require arbitrarily high prices. The value of lost load, VOLL, is the highest price that should be paid for power. It is what power is worth on average to the customers that would be curtailed by involuntary load shedding. If power is only worth \$10,000/MWh to these customers, there is no reason to pay \$20,000/MWh to provide them with power. In short, a price-cap at VOLL is always appropriate. If an unregulated market clears at higher prices, it is less efficient than one regulated with a VOLL price cap.

Another problem is both more common and more serious. In the current example, peak price approaches infinity as elasticity approaches zero. Because current power markets have some elasticity, this appears to indicate they should have a long-run equilibrium without a price cap. In this respect, the model is deceptive. Although current markets have some elasticity, it is limited in the sense that only a tiny fraction of demand is elastic. For example, if load at \$20/MWh is 8000MW, a \$1000 price increase might reduce it by 200MW. This does not indicate that demand can be reduced arbitrarily by further price increases. Two hundred megawatts may represent all of the available elasticity. Most customers do not pay attention to the hourly price and, even if they did, would receive no credit for responding to it. These loads are completely inelastic.

A market in which most load is entirely inelastic can fail to have any long-run equilibrium even though some load is elastic. The profit function, $SR_{\pi}(K)$, can help explain. There is some ICap level, K^* , which the inelastic portion of the (augmented) demand never exceeds. For higher values of ICap, the system is safe—demand can always be sufficiently contained by price to prevent load shedding. In economic terms, the supply and demand curves will always intersect. But, without a price limit imposed by regulators or courts, $SR_{\pi}(K)$ is extremely large (mathematically infinite) whenever ICap is less than K^* . If the market's demand elasticity is sufficiently small, then for all $K > K^*$, $SR_{\pi}(K) < FC$, but for $K \leq K^*$, supply will sometimes fail to intersect demand, price will rise without limit and average short-run profits will exceed fixed costs. Consequently there is no value of K for which $SR_{\pi} = FC$ [1 at 144, 152]. Depending on the shape of the load-duration curve and the elasticity of the demand curve, the market may or may not have a long-run equilibrium.

D. Interpretation of the Market Failure

The failure of the standard market mechanism to determine an equilibrium level of investment is not surprising. A preponderance of consumers do not respond to price. This is not a reflection of their desire to purchase power at any price, but a reflection of a structural problem with the market—lack of real-time billing. The information needed to determine the optimal level of reliability and installed capacity simply does not enter the market. Given this flaw, no market mechanism can ever determine that level.

Because consumers do not respond to price, those who buy power for consumers must decide how much to offer on their behalf. Typically, the system operator makes this decision, and it is the determining factor for investment. Under certain simplifying assumptions, if the system operator pays VOLL whenever load is shed and pays only the marginal cost at other times, the market will induce optimal investment. Even in this case, the market has not determined reliability. VOLL is not and, cannot be, determined by a market; it is set by regulators in order to achieve a certain level of reliability. In Australia, they have abandoned all attempts to determine its true value and simply set it to produce the “market” outcome that they want. On their own, present markets have no ability to determine optimal investment, and would produce unstable investment dynamics if left unregulated.

In fact, all U.S. markets have price limits and thus have long-run equilibria. In most markets, FERC has set a price cap, but where no cap has been formally prescribed, system operators set an informal cap as the need arises during emergencies. This is obviously not desirable, but such ad hoc decisions were documented during early Midwest price spikes. Such situations arise because the market-fundamentalist ideology has prevented rational discussion of price caps and thus prevented the formation of a coherent price-cap policy.

VI. THE ROLE OF ENGINEERING

A simplistic but useful view of the current situation is that FERC sets price caps, NERC sets operating reserve requirements and neither takes account of the other’s policy. Together these determine investment and thus long-run reliability. An equivalent view is that FERC determines the height of price spikes, NERC determines their duration, and the area (profit) is what matters.

Coordination is essential, but that will not happen until both regulators and engineers understand the tradeoff between height and duration—between price caps and operating-reserve requirements. (I practice this trade-off is complicated by unwritten policies regarding how the system operator raises price when the system is short of operating reserves.) This understanding would be most rapidly advanced by more-realistic calculations than those presented here. Engineers, can, with relatively little difficulty, pick a target level of installed capacity and compute the set of (P_{cap}, OR^R) pairs that will induce this level.

Engineers can also provide crucial information to regulators concerning the implications of these policy options for year-to-year profit variation. Using Monte Carlo simulations of weather and outages, they can compute a distribution of annual

short-run profits. Beside having implications for investment risk, a distribution that includes too many years of very-low, short-run profits will induce periods of low investment that lead to periodic lapses in reliability.

Reliability in a market environment requires an understanding of the market mechanism. That was missing in California and may have contributed to the reliability problems experienced there. Low prices in the early years may have contributed to underinvestment.

Engineers may also wish to argue for at least partial reliance on ICap requirements. These can reduce market volatility and ICap fluctuations. Thus, in a market with a crippled demand side they can increase investment efficiency. However, if they are combined with price-spikes they can lead to excess profits and overinvestment. To show that a particular combination does not lead to over-investment, both the ICap market and the energy market must be taken into account when computing short-run profits and designing the regulatory requirements. This is not a difficult engineering calculation.

In sum, there are only two logical choices for investment policy, (1) rely completely on the market and give up price caps, OpRes requirements and ICap requirements, or (2) compute the long-run impacts of these short-run policies and adjust them accordingly. The first choice courts disaster until the demand side has been fixed, and the present ostrich-like approach, which refuses to chose one or the other, is irresponsible.

VII. REFERENCES

- [1] S. Stoft, *Power System Economics: Designing Markets for Electricity*. New York: Wiley-IEEE Press, 2002.

VIII. BIOGRAPHY



Steven Stoft (1947) received a B.S. in Engineering Mathematics from the University of California at Berkeley where he focused on physics, programming, and digital circuits. After a year of graduate work in astrophysics at U.C. Santa Cruz, he taught electronics at a private high school and worked at the Exploratorium science museum in San Francisco. He then returned to U.C. Berkeley to earn a Ph.D.

in Economics.

After teaching at Boston University and U.C. Santa Cruz, he joined the Energy Analysis Program at the Lawrence Berkeley National Laboratory. He was also affiliated with the U.C. Energy Institute where he analyzed demand-side management and the restructuring of California’s electricity industry. He spent a year in the Office of Economic Policy at FERC which he left to start his own consulting practice. He is currently a consultant to PJM and the California Electricity Oversight Board. He has published numerous articles on markets for electricity, and is the author of the best-selling book on power market design: *Power System Economics*.