CSEM WP 129

Reliability and Competitive Electricity Markets*

Paul Joskow and Jean Tirole

April 2004

This paper is part of the Center for the Study of Energy Markets (CSEM) Working Paper Series. CSEM is a program of the University of California Energy Institute, a multi-campus research unit of the University of California located on the Berkeley campus.

2547 Channing Way
Berkeley, California 94720-5180
www.ucei.org

Reliability and Competitive Electricity Markets*

Paul Joskow† and Jean Tirole‡

April 21, 2004

Abstract

Despite all of the talk about "deregulation" of the electricity sector, a large number of non-market mechanisms have been imposed on emerging competitive wholesale and retail markets. These mechanisms include spot market price caps, operating reserve requirements, non-price rationing protocols, and administrative protocols for managing system emergencies. Many of these mechanisms have been carried over from the old regime of regulated monopoly and continue to be justified as necessary responses to market imperfections of various kinds and engineering requirements dictated by the special physical attributes of electric power networks. This paper seeks to bridge the gap between economists focused on designing competitive market mechanisms and engineers focused on the physical attributes and engineering requirements they perceive as being needed for operating a reliable electric power system. The paper starts by deriving the optimal prices and investment program when there are price-insensitive retail consumers, and their load serving entities can choose any level of rationing they prefer contingent on real time prices. It then examines the assumptions required for a competitive wholesale and retail market to achieve this optimal price and investment program. The paper analyses the implications of relaxing several of these assumptions. First, it analyzes the interrelationships between regulator-imposed price caps, capacity obligations, and system operator procurement, dispatch and compensation arrangements. It goes on to explore the implications of potential network collapses, the concomitant need for operating reserve requirements and whether market prices will provide incentives for investments consistent with these reserve requirements.

*We are grateful to Claude Crampes, Richard Green, Stephen Holland, Bruno Jullien, Patrick Rey and the participants at the IDEI-CEPR conference on "Competition and Coordination in the Electricity Industry," January 16–17, 2004, Toulouse and the ninth annual POWER conference, UC Berkeley, March 19, 2004, for helpful discussions and comments.

†Department of Economics, and Center for Energy and Environmental Policy Research, MIT.

‡IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT.

1 Introduction

Despite all of the talk about "deregulation" of the electricity sector, there continue to be a large number of non-market mechanisms that have been imposed on the emerging competitive wholesale and retail electricity markets. These mechanisms include: wholesale market price caps, capacity obligations placed on LSEs, frequency regulation, operating reserve and other ancillary service requirements enforced by the system operator, procurement obligations placed on system operators, protocols for non-price rationing of demand to respond to "shortages", and administrative protocols for system operators' management of system emergencies. Many of these non-market mechanisms have been carried over from the old regulated regime without much consideration of whether and how they might be replaced with market mechanisms and of the effects they may have on market behavior and performance if they are not.

In some cases the non-market mechanisms are argued to be justified by imperfections in the retail or wholesale markets: in particular, problems caused by the inability of most retail customers to see and react to real time prices with legacy meters, non-price rationing of demand, wholesale market power problems and imperfections in mechanisms adopted to mitigate these market power problems.

Other mechanisms and requirements have been justified by what are perceived to be special physical characteristics of electricity and electric power networks which in turn lead to market failures that are unique to electricity. These include the need to meet specific physical criteria governing network frequency, voltage and stability that are thought to have public good attributes, the rapid speed with which responses to unanticipated failures of generating and transmission equipment must be accomplished to continue to meet these physical network attributes and the possibility that market mechanisms cannot respond fast enough to achieve the

network's physical operating parameters under all states of nature.

Much of the economic analysis of the behavior and performance of wholesale and retail markets has either ignored these non-market mechanisms or failed to consider them in a comprehensive fashion. There continues to be a lack of adequate communication and understanding between economists focused on the design and evaluation of alternative market mechanisms and network engineers focused on the physical complexities of electric power networks and the constraints that these physical requirements may place on market mechanisms. The purpose of this paper and of Joskow-Tirole (2004) is to start to bridge this gap.

The institutional environment in which our analysis proceeds has competing load serving entities (LSEs)[1] that market electricity to residential, commercial and industrial ("retail") consumers. LSEs may be independent entities that purchase delivery services from unaffiliated transmission and distribution utilities or they may be affiliates of these transmission and distribution utilities that compete with unaffiliated LSEs. Some retail consumers served by LSEs respond to real time wholesale market prices, while others are on traditional meters which record only their total consumption over some period of time (for instance, a quarter), and therefore do not react to the real-time price. Retail consumers may be subject to non-price rationing to balance supply and demand in real time. The wholesale market is composed of competing generators who compete to sell power to LSEs. The wholesale market may be perfectly competitive or characterized by market power. Finally, there is an independent system operator (ISO) which is responsible for operating the transmission network in real time to support the wholesale and retail markets for power, including meeting certain network reliability and wholesale market power mitigation criteria.[2]

[1] Or in UK parlance "retail suppliers".

[2] The latter may include enforcing operating reserve and other operating reliability requirements, enforcing longer term capacity obligations, procuring and dispatching resources to meet these

Section 2 first derives the optimal prices and investment program when there is state contingent demand, at least some consumers do not react to real time prices, but their LSE can choose any level of rationing it prefers contingent on real time prices. In this model consumers are identical, possibly up to a proportionality factor, and therefore all have the same load profile. While the latter significantly constrains the nature of consumer heterogeneity considered, it is consistent with the existing literature (e.g., Borenstein-Holland, 2003). Joskow-Tirole (2004) analyzes more complex characterizations of consumer heterogeneity in the presence of retail competition. We then derive the competitive equilibrium under these assumptions when there are competing LSEs that can offer two part tariffs. This leads to a proposition that extends the standard, welfare theorem to price-insensitive consumers and rationing; this proposition serves as an important *benchmark* for evaluating a number of non-market obligations and regulatory mechanisms:

The second best optimum (given the presence of price-insensitive consumers) can be implemented by an equilibrium with retail and generation (wholesale) competition provided that:

(a) The real time wholesale price accurately reflects the social opportunity cost of generation.

(b) Rationing, if any, is orderly, and makes efficient use of available generation.

(c) LSEs face the real time wholesale price for the aggregate consumption of the retail customers for whom they are responsible.

(d) Consumers who can react fully to the real time price are not rationed. Furthermore, the LSEs serving consumers who cannot fully react to the real time price can demand any level of rationing they prefer contingent on the real-time price.

(e) Consumers have the same load profile (they are identical up to a scale factor).

The assumptions underlying this benchmark proposition are obviously very strong:

requirements, and managing system emergencies that might lead the network to collapse.

(a) market power on the one hand, and regulator-imposed price caps and other policy interventions on the other hand create differences between the real time wholesale market price and the social opportunity cost of generation; (b) network collapses, unlike say rolling blackouts, have systemic consequences, in that some available generation cannot be used to satisfy load; (c) LSEs do not face the real time price for their customers if these customers are load profiled; (d) price sensitive consumers may be rationed along with everyone else that is physically connected to the same controllable distribution circuit; and, relatedly, LSEs generally cannot demand any level of rationing they desire; (e) consumer heterogeneity is more complex than a scaling factor. This paper examines the implications of relaxing assumptions (a) and (b), while Joskow-Tirole (2004), that focuses on retail competition, investigates the failure of assumptions (c), (d), and (e).

Section 3 studies the implications of distorted wholesale prices. It first considers the case where there is a competitive supply of base load generation, market power in the supply of peak load investment and production, and a price cap is applied that constrains the wholesale market price to be lower than the competitive price during peak periods (section 3.1). This creates a shortage of peaking capacity in the long run when there is market power in the supply of peaking capacity. We show that capacity obligations and associated capacity prices have the potential to restore investment incentives by compensating generators ex ante for the shortfall in earnings that that they will incur due to the price cap. Indeed, with up to three states of nature (and up to two states of nature where generators have market power), the Ramsey optimum can be achieved despite the presence of market power through a combination of a price cap and capacity obligations provided that : (i) both peak and base load generating capacity are eligible to meet LSE capacity obligations and receive the associated capacity price, and (ii) the demand of all consumers, including price-sensitive consumers, counts for determining capacity obligations and

the capacity prices are reflected in the prices paid by all retail consumers. With more than three states of nature, a combination of spot wholesale market price caps and capacity obligations will not achieve the Ramsey optimum unless market power is only a problem during peak demand periods. Thus, the regulator faces a tradeoff between alleviating market power off-peak, if it is a problem, through a strict price cap, and providing the proper peak investment incentives, and is further unable to provide price-sensitive consumers with the appropriate economic signals. The intuition for this result is that when more than two prices are distorted by market power the optimality of a competitive equilibrium cannot be restored with only two instruments — a price cap and a capacity obligation.

Section 3.2 then examines the effects of two types of behavior by an ISO that empirical analysis has suggested may distort prices and investment (Patton 2002). The first involves inefficient or "out-of-merit" dispatch of resources procured by the ISO. Such dispatch in the short run depresses off-peak prices and in the long term leads to an inefficient substitution of base load units by peakers. The second involves the recovery of the costs of resources acquired by the ISO through an uplift charge spread over prices in all demand states or else in only peak demand states. Whether the uplift is socialized (spread over demand states) or not, large ISO purchases discourage the build up of baseload capacity and depresses the peak price. For small purchases, off-peak capacity decreases under a socialized uplift, and peak capacity decreases under an uplift that applies solely to peak energy consumption.

Section 4 derives the implications of network collapses and the concomitant need for network support services. As suggested above, network collapses differ from other forms of energy shortages and rationing in a fundamental way. While scarcity makes available generation (extremely) valuable under orderly rationing, it makes it valueless when the network collapses.[3] Hence, system collapses, unlike, say, controlled

[3]An analogy may help understand the distinction between orderly rationing and a collapse:

rolling blackouts that shed load to match demand with available capacity, create a rationale for network support services with public goods characteristics. We derive the optimal level for these network support services, and discuss the implementation of the Ramsey allocation through a combination of operating reserve obligations and market mechanisms.

2 A benchmark decentralization result

2.1 Model[4]

There is a continuum of states of nature or periods $i \in [0,1]$. The frequency of state i is denoted f_i (and so $\int_0^1 f_i di = 1$). Let $E[\cdot]$ denote the expectation operator with respect to the density f_i.[5] We assume that the (unrationed) demand functions of price-insensitive and price-sensitive consumers, D_i and $\widehat{D}_i$, are increasing in i.[6]

Price-insensitive consumers are on traditional meters that record only their aggregate consumption over all states of nature, and therefore do not react to the RTP.[7] Consumers are homogeneous, up to possibly a scaling factor, i.e., they have the same load profile. [Note that if consumers differ in the scale of their demand, this scale can be inferred from total consumption and need not be known by the social planner or the LSEs.] Without loss of generality they are offered a two-part tariff, with a fixed fee A and a marginal price p. Their demand function in the absence

when a mattress manufacturer fails, buyers of mattresses may experience delays; competitors however do not suffer and may even gain from the failure. By contrast, a farmer whose cows have contracted the mad cow disease may spoil the entire market for beef.

[4]See Turvey and Anderson (1977, Chapter 14) for an analysis of peak period pricing and investment under uncertainty when prices are fixed ex ante and all demand is subject to rationing with a constant cost of unserved energy when demand exceeds available capacity.

[5]$E[x_i] = \int_0^1 x_i f_i di$.

[6]In this paper, we do not allow intertemporal transfers in demand (demand in state i depends only on the price faced by the consumer in state i). We could allow such transfers, at the cost of increased notational complexity.

[7]As in Joskow-Tirole (2004), we could also introduce consumers on real-time meters who do not monitor the real-time price. This would not affect Proposition 1 below.

of rationing is denoted $D_i(p)$, with D_i increasing in i. We let $\alpha_i \leq 1$ denote the fraction of their demand satisfied in state i. As α_i decreases, the fraction of load interrupted $(1-\alpha_i)$ increases. The alphas may be exogenous (say, determined by the system operator); alternatively, one could envision situations in which the LSEs would affect the alphas either by demanding that their consumers not be served as the wholesale price reaches a certain level, or conversely by bidding for priority in situations of rationing.[8] We let $\mathcal{D}_i(p, \alpha_i)$ denote their expected consumption in that state, and $\mathcal{S}_i(p, \alpha_i)$ their realized gross surplus, with

$$\mathcal{D}_i(p, 1) = D_i(p) \quad \text{and} \quad \mathcal{S}_i(p, 1) = S_i(D_i(p)),$$

where S_i is the standard gross surplus function (with $S_i' = p$). We assume that $\mathcal{S}_i$ is concave in α_i on $[0, 1]$: more severe rationing involves higher relative deadweight losses.

In the *separable case*, the demand $\mathcal{D}_i$ takes the multiplicative form $\alpha_i D_i(p)$ and the surplus takes the separable form $\mathcal{S}_i(D_i(p), \alpha_i)$. More generally however, the consumer may adjust her demand to the prospect of being potentially rationed.[9]

We will also assume that lost opportunities to consume do not create value to the consumer. Namely, the net surplus

$$\mathcal{S}_i(p, \alpha_i) - p\mathcal{D}_i(p, \alpha_i)$$

is maximized at $\alpha_i = 1$, that is, when it is equal to $S_i(D_i(p)) - pD_i(p)$.

Let us now discuss specific cases to make this abstract formalism more concrete, and note that the social cost of shortages depends on how fast demand and supply

[8] The latter of course assumes that the system operator can discriminate in its dispatch to LSEs in each state, including in emergency situations that require the system operator to act quickly to avoid a cascading blackout.

[9] A case in point is voltage reduction. When the system operator reduces voltage by, say, 5%, lights become dimmer, motors run at a slower pace, and so on. A prolonged voltage reduction, though, triggers a response: consumers turn on more lights, motor speeds are adjusted. Another example of non-separability will be provided below.

conditions change relative to the reactivity of consumers.[10]

When the timing of the blackout is perfectly anticipated and blackouts are rolling across geographical areas, then α_i denotes the population percentage of geographical areas that are not blacked out (and thus getting full surplus $S_i(D_i(p))$), and $1-\alpha_i$ the fraction of consumers living in dark areas (and thus getting no surplus from electricity). With perfectly anticipated blackouts, it makes sense to assume that

$$\mathcal{S}_i(p,\alpha_i) = \alpha_i S_i(D_i(p)) \quad \text{and} \quad \mathcal{D}_i(p,\alpha_i) = \alpha_i D_i(p).$$

An unexpected blackout may have worse consequences than a planned cessation of consumption. For example, a consumer may prefer using the elevator to the stairs. If the outage is foreseen, then the consumer takes the stairs (does not "consume" the elevator) and gets zero surplus from the elevator. By contrast, the consumer obtains a negative surplus from the elevator if the outage is unforeseen. Similarly, consumers would have planned an activity requiring no use of electricity (going to the beach rather than using the washing machine, drive their car or ride their bicycle rather than use the subway) if they had anticipated the blackout; workers could have planned time off, etc. More generally, with adequate warning consumers can take advance actions to adapt to the consequences of an interruption in electricity supplies. This is one reason why distribution companies notify consumers about planned outages required for maintenance of distribution equipment.

Opportunity cost example: Suppose that the consumer chooses between an electricity-consuming activity (taking the elevator, using electricity to run an equipment) and an electricity-free approach (taking the stairs, using gas to run the equipment). The latter yields known surplus $\overline{S} > 0$. The surplus associated with the former depends not only on the marginal price p he faces for electricity, but also on the probability $1-\alpha_i$ of not being served. One can envision three information structures: (a)

[10] This observation is made for example in EdF (1994, 1995).

The consumer knows whether he will be served (the elevator is always deactivated through communication just before the outage); this is the foreseen rolling blackouts case just described. (b) The consumer knows the state-contingent probability α_i of being served, but he faces uncertainty about whether the outage will actually occur (he knows that the period is a peak one and he is more likely to get stuck in the elevator). (c) The consumer has no information about the probability of outage and bases his decision on $E[\alpha_i]$ (he just knows the average occurrence of immobilizations in elevators). Letting $S_i^n(p) \equiv \max\{S_i(D) - pD\}$ denote the net surplus in the absence of rationing; then

$$\mathcal{S}_i(p,\alpha_i) - p\mathcal{D}_i(p,\alpha_i) = \begin{cases} \alpha_i S_i^n(p) + (1-\alpha_i)\overline{S} & \text{in case (a)} \\ \max\left\{\alpha_i S_i^n(p), \overline{S}\right\} & \text{in case (b)} \\ \alpha_i S_i^n(p) & \text{in case (c)} \end{cases}$$

(provided that $S_i^n(p) \geq \overline{S}$ and, in case (c), that $E[\alpha_i]$ is high enough so that the consumer chooses the electricity-intensive approach).

The value of lost load (VOLL) is equal to the marginal surplus associated with a unit increase in supply to these consumers, and is here given by

$$\text{VOLL}_i = \frac{\dfrac{\partial \mathcal{S}_i}{\partial \alpha_i}}{\dfrac{\partial \mathcal{D}_i}{\partial \alpha_i}},$$

since a unit increase in supply allows an increase in α_i equal to $1/[\partial \mathcal{D}_i/\partial \alpha_i]$. When $\mathcal{D}_i = \alpha_i D_i$, then

$$\text{VOLL}_i = \frac{\dfrac{\partial \mathcal{S}_i}{\partial \alpha_i}}{D_i}.$$

For example, with perfectly anticipated blackouts, the value of lost load is equal to the average gross consumer surplus. It is higher for unanticipated blackouts than for blackouts that give consumers time to adapt their behavior in anticipation of

being curtailed.

Price-sensitive consumers are modeled in exactly the same way and obey the exact same assumptions as price-insensitive consumers. The only difference is that they face the real time price and react to it. Let $\widehat{p}_i$ denote the state-contingent price chosen by the social planner; although we will later show that it is optimal to let price-sensitive consumers face the RTP p_i (so $\widehat{p}_i = p_i$), we must at this stage allow the central planner to introduce a wedge between the two prices. In state i their expected consumption is $\widehat{\mathcal{D}}_i\left(\widehat{p}_i, \widehat{\alpha}_i\right)$ and their gross surplus is $\widehat{\mathcal{S}}_i\left(\widehat{p}_i, \widehat{\alpha}_i\right)$, where $\widehat{\alpha}_i$ is the rationing / interruptibility factor for price-sensitive consumers.

The supply side is described as a continuum of investment opportunities indexed by the marginal cost of production c. Let $I(c)$ denote the investment cost of a plant producing one unit of electricity at marginal cost c. There are constant returns to scale for each technology. We denote by $G(c) \geq 0$ the cumulative distribution function of plants.[11] So, the total investment cost is

$$\int_0^\infty I(c) dG(c).$$

The ex post production cost is

$$\int_0^\infty c u_i(c) dG(c), \quad \text{where} \quad \int_0^\infty u_i(c) dG(c) = Q_i.$$

where the utilisation rate $u_i(c)$ belongs to $[0, 1]$.

Remark: The uncertainty is here generated on the demand side. We could add an availability factor λ (a fraction $\lambda \in [0, 1]$ of plants is available, where λ is given by some cdf $H_i(\lambda)$) as in Section 4 below. This would not alter the conclusions.

[11] This distribution may not admit a continuous density. For example, only a discrete set of equipments may be selected at the optimum.

2.2 Optimum

A social planner chooses a marginal price p for price-insensitive consumers, and (for each state i) marginal prices $\widehat{p}_i$ for price-sensitive consumers, the extents of rationing α_i and $\widehat{\alpha}_i$, utilisation rates $u_i(\cdot)$ and the investment plan $G(\cdot)$ so as to solve:

$$\max\left\{E\left[\mathcal{S}_i\left(p,\alpha_i\right)+\widehat{\mathcal{S}}_i\left(\widehat{p}_i,\widehat{\alpha}_i\right)-\int_0^\infty u_i(c)cdG(c)\right]-\int_0^\infty I(c)dG(c)\right\}$$

s.t.

$$\int_0^\infty u_i(c)dG(c)\geq\mathcal{D}_i\left(p,\alpha_i\right)+\widehat{\mathcal{D}}_i\left(\widehat{p}_i,\widehat{\alpha}_i\right)\quad\text{for all } i.$$

Letting $p_i f_i di$ denote the multiplier of the resource constraint in state i, the first-order conditions yield:

a) *Efficient dispatching*:

$$u_i(c)=1\quad\text{for}\quad c<p_i\quad\text{and}\quad u_i(c)=0\quad\text{for}\quad c>p_i. \tag{1}$$

b) *Price-sensitive consumers*:[12]

$$\begin{aligned}&\text{(i)}\qquad\widehat{\mathcal{D}}_i=\widehat{D}_i\left(p_i\right)\\&\text{(ii)}\qquad\widehat{\alpha}_i=1.\end{aligned} \tag{2}$$

[12]To prove condition (2), apply first the observation that by definition $\mathcal{D}_i\left(p_i,\widehat{\alpha}_i\right)$ is the net-surplus-maximizing quantity for a consumer paying price p_i for a given probability $\widehat{\alpha}_i$ of being served; and second our assumption that lost opportunities don't create value:

$$\begin{aligned}\widehat{\mathcal{S}}_i\left(\widehat{p}_i,\widehat{\alpha}_i\right)-p_i\widehat{\mathcal{D}}_i\left(\widehat{p}_i,\widehat{\alpha}_i\right)&\leq\widehat{\mathcal{S}}_i\left(p_i,\widehat{\alpha}_i\right)-p_i\widehat{\mathcal{D}}_i\left(p_i,\widehat{\alpha}_i\right)\\&\leq\widehat{S}_i\left(\widehat{D}_i\left(p_i\right)\right)-p_i\widehat{D}_i\left(p_i\right).\end{aligned}$$

Hence, price sensitive consumers should not be rationed and should face price p_i.

c) *Price-insensitive consumers*:

$$\text{(i)} \qquad E\left[\frac{\partial \mathcal{S}_i}{\partial p} - p_i \frac{\partial \mathcal{D}_i}{\partial p}\right] = 0.$$

$$\text{(ii)} \qquad \text{Either} \quad \frac{\dfrac{\partial \mathcal{S}_i}{\partial \alpha_i}}{\dfrac{\partial \mathcal{D}_i}{\partial \alpha_i}} = p_i \quad \text{or} \quad \alpha_i = 1. \tag{3}$$

d) *Investment*:

$$\text{Either} \qquad I(c) = E\left[\max\left\{p_i - c, 0\right\}\right] \quad \text{or} \quad dG(c) = 0. \tag{4}$$

These first-order conditions can be interpreted in the following way: condition (1) says that only plants whose marginal cost is smaller than the dual price p_i are dispatched in state i. Condition (2) implies that price-sensitive consumers are never rationed and that their consumption decisions are guided by the state-contingent dual price. Condition (3) yields the following formula for the price $p = p^*$ provided that price-insensitive consumers are never rationed ($\alpha_i \equiv 1$):

$$E\left[\left(p^* - p_i\right) D_i'\left(p^*\right)\right] = 0. \tag{5}$$

In case of rationing ($\alpha_i < 1$ for some i), its implications depend on the efficiency of rationing; condition (3) in the *separable case* yields the following formula:

$$E\left[\left[\frac{\partial \mathcal{S}_i}{\partial D_i} - \alpha_i p_i\right] D_i'(p)\right] = 0.$$

For example, for *perfectly foreseen outages*, it boils down to:[13]

$$E\left[\left(p - p_i\right)\left[\alpha_i D_i'\left(p\right)\right]\right] = 0. \tag{6}$$

[13]Suppose that the regulator imposes an artificial constraint that retail customers not be voluntarily shut off (one may have in mind a small fraction of such customers, so that the wholesale prices is not affected). The Ramsey price would then be p^*. Under the reasonable assumption that α_i decreases and $(p_i - p)\left|D_i'\right|$ increases with the state of nature and decreases with p, (6) yields a corrected Ramsey price p^{**}:

$$p^{**} < p^*.$$

Intuitively, the impact of p on peak demand is reduced by rationing, and so there is less reason to keep the marginal price high.

Condition (3ii) implies that in all cases of rationing

$$\text{VOLL}_i = p_i.$$

That is, generators and LSEs should all face the value of lost load.

Finally, condition (4) is the standard free-entry condition for investment in generation.

2.3 Competitive equilibrium

Let us now assume that price-sensitive and -insensitive consumers are served by load serving entities (LSEs), and that LSEs face the real time wholesale price for the aggregate consumption of the retail customers for whom they are responsible.

The following proposition shows that, despite rationing and price-insensitive consumptions, retail competition is consistent with Ramsey optimality provided that five assumptions are satisfied:

Proposition 1 *The second-best optimum (that is, the socially optimal allocation given the existence of price-insensitive retail consumers) can be implemented by an equilibrium with retail and generation competition provided that:*

- *the RTP reflects the social opportunity cost of generation,*
- *available generation is made use of during rationing periods,*
- *load-serving entities face the RTP,*

By contrast, with imperfectly foreseen outages,

$$\frac{\partial \mathcal{S}_i}{\partial D_i} > \alpha_i p,$$

and (3) yields a price above p^{**}. The increase in outage cost due to unforeseeability suggests raising the marginal price to retail consumers in order to suppress demand.

- *price-sensitive consumers are not rationed; furthermore, while price-insensitive consumers may be rationed, their load-serving entity can demand any level of state-contingent rationing* $\alpha_i(p_i)$,[14]

- *consumers are homogeneous (possibly up to a scaling factor).*

Proof: Suppose that competing retailers (LSEs) can offer to price-insensitive contracts $\{A, p, \alpha.\}$, that is two-part tariffs with fixed fee A and marginal price p cum a state-contingent extent of rationing α_i. Retail competition induces the maximization of the joint surplus of the retailer and the consumer:

$$\max_{\{p,\alpha.\}} E\left[\mathcal{S}_i(p,\alpha_i) - p_i\mathcal{D}_i(p,\alpha_i)\right].$$

The first-order conditions for this program are nothing but conditions (3) above. The rest of the economy is standard, and so the fundamental theorem of welfare economics applies. ■

The assumptions underlying Proposition 1 are very strong: In practice, (a) market power on the one hand, and price caps and other policy interventions on the other hand create departures of RTPs from the social opportunity cost of generation; and (b) available generation does not serve load during blackouts associated with a network collapse; (c) LSEs do not face the RTP for the power they purchase in the wholesale market if their customers are load profiled; (d) technological constraints in the distribution network imply that price-sensitive consumers may be rationed along with everyone else; relatedly, LSEs cannot generally demand any level of rationing they desire; (e) consumer heterogeneity is more complex than a simple scaling factor. The paper investigates the consequences of the first two observations.

Remark: Chao-Wilson (1987) also emphasize the use of bids for priority servicing.

[14]Here the state and the price are mapped one-to-one. More generally, they may not be (the state of nature involves unavailability of plants, say). The proposition still holds as long as LSEs can select a state-contingent α_i.

Chao and Wilson show that when consumers are heterogeneous and have unit and state-independent demands, the first-best (hence rationing free) allocation can be implemented equivalently through a spot market or an ex ante priority servicing auction. Proposition 1 by contrast considers homogeneous consumers and introduces price insensitive consumers; accordingly, markets are here only second-best optimal and the second-best optimal allocation involves actual rationing.

2.4 Two-state example

There are two states: off-peak ($i = 1$) and peak ($i = 2$), with frequencies f_1 and f_2 ($f_1 + f_2 = 1$); price insensitive retail customers have demands $D_1(p)$ and $D_2(p)$ with associated gross surpluses (in the absence of rationing) $S_1\left(D_1(p)\right)$ and $S_2\left(D_2(p)\right)$. Price-sensitive customers (who react to real-time pricing) have demands $\widehat{D}_1(p)$ and $\widehat{D}_2(p)$, with associated gross surpluses (in the absence of rationing) $\widehat{S}_1\left(\widehat{D}_1(p)\right)$ and $\widehat{S}_2\left(\widehat{D}_2(p)\right)$. We assume that rationing may occur only at peak ($\alpha_1 = 1\ ,\ \alpha_2 \leq 1$).

A unit of baseload capacity costs I_1 and allows production at marginal cost c_1. Let K_1 denote the baseload capacity. The unit cost of installing peaking capacity is I_2. The marginal operating cost of the peakers is c_2.

Social optimum: Letting p^* denote the (constant) price faced by retail consumers, the (second-best) social optimal solves over $\left\{p^*, \alpha_2\,, \widehat{D}_1, \widehat{D}_2\right\}$

$$\begin{aligned}\max W = \max \Big\{ & f_1\left[S_1\left(D_1\left(p^*\right)\right) + \widehat{S}_1\left(\widehat{D}_1\right) - c_1 K_1\right] - I_1 K_1 \\ & + f_2\left[\mathcal{S}_2\left(p^*, \alpha_2\right) + \widehat{S}_2\left(\widehat{D}_2\right) - c_1 K_1 - c_2 K_2\right] - I_2 K_2\Big\}\end{aligned}$$

where

$$K_1 \equiv D_1\left(p^*\right) + \widehat{D}_1 \tag{7}$$

$$K_2 \equiv \left[\mathcal{D}_2\left(p^*, \alpha_2\right) + \widehat{D}_2\right] - \left[D_1\left(p^*\right) + \widehat{D}_1\right] \tag{8}$$

Applying the general analysis yields (provided that the peakers' marginal cost c_2 weakly exceeds the off-peak price p_1):

$$\text{Either } \widehat{S}_i' = p_i \text{ or } \widehat{D}_i = 0, \tag{2'i}$$

$$f_1 (p^* - p_1) D_1' + f_2 \left(\frac{\partial \mathcal{S}_2}{\partial p} - p_2 \frac{\partial \mathcal{D}_2}{\partial p} \right) = 0, \tag{3'i}$$

and

$$\begin{aligned} & f_1 (p_1 - c_1) + f_2 (p_2 - c_1) = I_1 \\ & f_2 (p_2 - c_2) = I_2. \end{aligned} \tag{4'i}$$

Note that the free entry investment conditions imply that the peak price exceeds the marginal operating cost of peaking capacity in equilibrium.

Proposition 2 *Rationing ($\alpha_2 < 1$) of price-insensitive consumers may be optimal.*

Proof: With foreseen rolling blackouts, $\mathcal{S}_2 (p^*, \alpha_2) = \alpha_2 S_2 (D_2 (p^*))$ and so rationing is desirable if and only if $S_2 (D_2 (p^*)) < p_2 D_2 (p^*)$, that is intuitively when the peak price is high. Suppose for example that f_2 is small (*infrequent peak*); then from (4') $f_2 p_2 \simeq I_2$, and $p_1 - c_1 \simeq I_1 - I_2$. If furthermore demand is linear and $D_1' = D_2'$, and $\alpha_2 = 1$, $p^* \simeq p_1 + I_2 = I_1 + c_1$ from (3'i). So p^* remains bounded, and $S_2 (D_2 (p^*))$ is indeed smaller than $p_2 D_2 (p^*)$, so rationing is optimal. ■

Intuitively, rationing serves to re-create some price sensitivity of the consumption of consumers on traditional meters. For example, for infrequent peaks, the peak price goes to infinity and so the discrepancy between the true price and the price paid by retail consumers is too large to make it socially optimal to serve these consumers.

3 Price distortions: capacity obligations and ISO procurement

3.1 Price caps and capacity obligations

A capacity obligation requires an LSE to contract for enough capacity to meet its peak demand (plus a reserve margin in a world with uncertain equipment outages and demand fluctuations). Capacity obligations may take at least two forms. One requires LSEs to forward contract with generators to make their capacity available to the ISO during peak demand periods, leaving the price for any energy supplied by this capacity[15] to be determined ex post in the spot market. Alternatively, the capacity obligations could require forward contracting for both capacity and the price of any energy (or operating reserves) supplied by that capacity during peak hours.[16]

Proposition 1 shows that rationing alone does not create a rationale for capacity obligations. Rather, there must be some reason why the spot price does not fully adjust to reflect supply and demand conditions and differs from the correct economic signal. Leaving aside procurement by the ISO for the moment, we can look in three directions. For this purpose, and like in section 2.4, we specialize the model in most of this section to *two states of nature*.

Market power in the wholesale market

The regulator may impose a price cap ($p_2 \leq p^{\max}$) on wholesale power prices, which in turn are reflected directly in retail prices given perfect competition among

[15] Or in a world with uncertain equipment outages and demand fluctuations the prices for operating reserves provided by this capacity as well.

[16] Another approach is for the system operator to purchase reliability contracts from generators on behalf of the load. Vazquez et al (2001) have designed a more sophisticated capacity obligations scheme, in which the system operator purchases reliability contracts that are a combination of a financial call option with a high predetermined strike price and an explicit penalty for non-delivery. Such capacity obligations are bundled with a hedging instrument, as the consumer purchasing such a call option receives the difference between the spot price and the strike price whenever the former exceeds the latter.

retailers, in order to prevent generators from exercising market power in the wholesale market during peak demand periods.

Suppose that:

- baseload investment and production is competitive (as earlier),
- peakload investment and production are supplied by an n-firm Cournot oligopoly.

We have in mind a relatively short horizon (certainly below 3 years), so that new peaking investment cannot be built in response to strategic withholding (in this interpretation, I_2 is probably best viewed as the cost of maintaining existing peakers). The timing has two stages: First, firms choose the capacity that they will make available to the market. Second, they supply this capacity in the market for energy. We leave aside rationing for simplicity.

In the absence of a price cap, an oligopolist in the peaking capacity market chooses the amount of capacity to make available to the market K_2^i so as to solve:

$$\max_{p_2} \left\{ \left[f_2 \left(p_2 - c_2 \right) - I_2 \right] \left[D_2 \left(p \right) + \widehat{D}_2 \left(p_2 \right) - K_1 - \sum_{j \neq i} K_2^j \right] \right\}.$$

Letting $\widehat{\eta}_2 \equiv -\dfrac{\partial \widehat{D}_2}{\partial p_2} / \dfrac{\widehat{D}_2}{p_2}$ denote the elasticity of demand of the price sensitive customers, one obtains the following Lerner formula:[17]

$$\frac{p_2 - \left[c_2 + \dfrac{I_2}{f_2} \right]}{p_2} = \frac{1}{n \widehat{\eta}_2} \left[\frac{\left[\widehat{D}_2 \left(p_2 \right) - \widehat{D}_1 \left(p_1 \right) \right] + \left[D_2(p) - D_1(p) \right]}{\widehat{D}_2 \left(p_2 \right)} \right] > 0, \qquad (9)$$

or,

$p_2 = p_2^C$, where p_2^C is the Cournot price.

[17] The other equilibrium conditions are:

$$K_1 = D_1(p) + \widehat{D}_1(p_1),$$
$$f_1 p_1 + f_2 p_2 = c_1 + I_1,$$

and

$$E\left[\left(p - p_i \right) D_i'(p) \right] = 0.$$

As expected, the *oligopolistic relative markup decreases with the number of firms and with the elasticity of demand of price-sensitive consumers, and decreases when price-insensitive consumers become price-sensitive.*[18]

A price cap $p^{\max} = p_2^* = c_2 + (I_2/f_2)$ restores the Ramsey optimum. By contrast, a price cap creates a shortage of peakers whenever $p^{\max} < p_2^*$.[19]

Let us now show that (i) with two states of nature, the Ramsey optimum can nevertheless be attained through capacity obligations even if the price cap is set too low, and (ii) with three states of nature, the combination of a price cap and capacity obligations restores the Ramsey optimum provided that the price cap is set at the competitive level in the lowest-demand state in which there is market power.

With two states of nature and a price cap that is set too low, to get the same level of investment and production in the second best as in the competitive equilibrium, the oligopolists must receive a capacity price p_K satisfying

$$I_2 - p_K = f_2 \left(p^{\max} - c_2\right).$$

[We assume that, as in PJM, the firm must supply K_2 ex post if requested to do so, and so ex post withholding of supplies is not an issue.]

Note that

$$p_K + f_2 p^{\max} = f_2 p_2^*$$

and so

$$I_1 - p_K = f_2 \left(p^{\max} - c_1\right) + f_1 \left(p_1 - c_1\right),$$

[18] Through the installation of a communication system, say. Because price-sensitivity reduces the consumption differential between peak and off peak, the numerator on the right-hand side of (9) decreases (and $\widehat{D}_2$ increases) as some more consumers become price-sensitive.

[19] The simple two-state example analyzed here assumes that during peak periods the price cap has been set below p_2^* to characterize the more general case in which the price cap is, on average, lower than the competitive market price. If the price cap were set high enough to ensure that $p^{\max} = p_2^*$ it would not lead to shortages of peaking capacity. However, the \$1000/MWh (or lower) price caps that are now used in the U.S. appear to us to be significantly lower than the VOLL in some high demand states.

so incentives for baseload production are unchanged, *provided that off-peak plants are made eligible for capacity payments.*[20]

There are at least four potential problems that may result from a policy of applying binding price caps to the price of energy sold in the wholesale spot market:

• *The price-sensitive customers then consume too much*: They consume $\widehat{D}_2\left(p^{\max}\right)$ at peak. The price paid by all retail consumers must also include the price of capacity p_K in order to restore proper incentives on the demand side.

• *The signal for penalizing a failure to deliver is lost*: The ISO no longer has a measure of the social cost associated with a supplier's failure to deliver ($p^{\max}$ is an underestimate of this cost). Similarly, there is no objective penalty for those LSEs that underpredict their peak demand and are short of capacity obligations.[21]

• *Ex ante monopoly behavior*: If one just lets the oligopolists choose the number of capacity contracts q_2^i, then the oligopolists are likely to restrict the number of these contracts. Actually, in the absence of price-insensitive consumers and assuming that the number of generators is the same (n) in the capacity and wholesale spot markets,[22] one can show a *neutrality result*: The outcome with ex post price cap and ex ante capacity obligation is the same as that with no price cap and no capacity obligation. The oligopolists just exploit their monopoly power ex ante.

To see this, note that consumers must pay $\frac{p_K}{f_2} + p^{\max}$ per unit of peak consumption. Oligopolist i therefore chooses to offer an amount q_2^i of capacity contracts

[20]Note that in New England, New York and PJM, all generating capacity meeting certain reliability criteria counts as ICAP capacity and can receive ICAP payments.

[21]In either case, there are then more than two states or nature (but see below the remark on idiosyncratic shocks).

[22]The ex ante market might be more competitive than the ex post market, in which capacity constraints are binding (this is the view taken for example in Chao-Wilson 2003). If so, how much more competitive depends on the horizon. Competition in peaking generation may be more intense 3 years ahead than 6 months ahead, and a fortiori a day ahead.

solving

$$\max_{p_K} \left\{ \left[f_2 \left(p^{\max} - c_2 \right) + p_K - I_2 \right] \left[\widehat{D}_2 \left(\frac{p_K}{f_2} + p^{\max} \right) - K_1 - \sum_{j \neq i} q_2^j \right] \right\}.$$

The first-order condition is the same as (9), with

$$\frac{p_K}{f_2} + p^{\max} = p_2^C.$$

The analysis with price-insensitive consumers is more complex, because the oligopolists can through the capacity market affect the price p offered by LSEs to price-insensitive consumers and thereby the latter's peak consumption, while they took $D_2(p)$ as given in our earlier analysis of spot markets.

An issue involving the nature of the contract supporting the capacity obligation has become somewhat confused in the policy discussions about capacity obligations. If the contract establishes an ex ante price for the right to call on a specified quantity of generating capacity in the future but the price for the energy to be supplied ex post is not specified in the forward contract, then, as shown above, the contracts supporting the capacity obligation are unlikely to be effective in mitigating market power unless the market for such contracts is more competitive than the spot market. If the capacity obligation is met with a contract that specifies both the capacity price and the energy supply price ex ante then such forward contracts can mitigate market power even if the forward market is no more competitive than the spot market.

It is well known that when generators have forward contract positions that specify the price at which they are committed to sell electricity their incentives to exercise market power in the spot market are reduced (Wolak 2000, Green 1999). For example, if a generator has contracted forward to sell all of its capacity at a fixed price p_f in each hour for the next three years it receives no benefit from withholding output from the spot market to drive up prices to a level greater than p_f. Indeed, in this case withholding output to drive up prices would reduce the generator's profit

since it would now have to buy enough power to make up for the supplies from the capacity it withheld at an inflated price.

A more controversial issue is whether and under what conditions (risk sharing considerations aside) a generator with market power in the spot market would enter into forward contracts with an overall price level lower than what they could expect to realize by not engaging in forward contracting and exercising market power in the spot market. That is, why aren't the benefits of any market power generators expect to realize in the spot market reflected in the forward contract prices they would agree to sign voluntarily as well? There has been a considerable amount of theoretical research that supports the view that except in the monopoly case forward markets will be more competitive than spot markets for electricity. Relevant papers include Allaz (1992), Allaz-Vila (1993), Green (1999), Newbery (1998), and Chao-Wilson (2003).

• *A capacity payment is an insufficient instrument with more than three states of nature.* The capacity payment p_K should compensate for the revenue shortfall (relative to the socially optimal price) created by the price cap *at peak*. With many states of nature and many means of production (as in section 2.2), the capacity payment can still compensate for the expected revenue shortfall for peakers and therefore for non-peakers as well if the price cap corrects for market power at peak. However, the price cap then fails to properly correct market power just below peak. Conversely, a price cap can correct for an arbitrary number of periods/ state of nature in which there is market power, provided that the plants be dispatchable in order to qualify for capacity obligations;[23] but, it then fails to ensure cost recovery for the peakers. To see this, suppose that $i \in [0, 1]$ as earlier, and that there is market power for

[23]The dispatching requirement comes from the fact that (with more than three states) the price cap may need to be lower than the marginal cost of some units that are dispatched in the Ramsey optimum. Also, note that the ISO must be able to rank-order plants by marginal cost in order to avoid inefficient dispatching.

$i \geq i_0$. The price cap must be set so that:

$$p^{max} = p^*_{i_0}.$$

Cost recovery for plants that in the Ramsey optimum operate if and only if $i \geq i_0$ requires that:

$$p_K = E\left[(p^*_i - p^{max})\, \mathbb{1}_i \geq i_0\right]$$

(where $\mathbb{1}_i \geq i_0 = 1$ if $i \geq i_0$ and 0 otherwise). But then a higher marginal cost plant, that should operate when $i \geq k > i_0$ *over*-recoups its investment as:

$$p_K > E\left[(p^*_i - p^{max})\, \mathbb{1}_i \geq k\right].$$

Similarly, the combination of a price cap and a capacity payment cannot provide the proper signals in all states of nature to price-sensitive consumers, if there are more than three states. With three states ($i = 1, 2, 3$), though, the price cap can be set at p^*_2. Then $f_3\left(p^*_3 - p^{max}\right) = p_K$ implies that $f_2\left(p^*_2 - p^{max}\right) + f_3\left(p^*_3 - p^{max}\right) = p_K$.

This reasoning has a standard "instruments vs targets" flavor. When more than two prices are distorted by market power (which, incidentally, also would have been the case with three states of nature, had we assumed that none of the markets was competitive), two instruments, namely a price cap and a capacity price, cannot restore optimality of the competitive equilibrium.

Remark: We have considered only aggregate uncertainty. However, a price-sensitive industrial consumer (or an undiversified LSE) further faces *idiosyncratic* uncertainty. A potential issue then is that while the capacity payment can supply the consumer with a proper *average* incentive to consume during peak (say, when there are two aggregate states), it implies that the consumer will overconsume for low idiosyncratic demand (as she faces a "low" price p^{max} at the margin) and underconsumes in high states of idiosyncratic demand (provided that penalties for exceeding the capacity obligation are stiff). This problem can however be avoided, provided that consumers

regroup to iron out idiosyncratic shocks (in a mechanism similar to that of "bubbles" in emission trading programs, or to the reserve sharing arrangements that existed prior to the restructuring of electricity systems).[24]

Proposition 3 *Capacity obligations have the potential to restore investment incentives by compensating generators ex ante for the shortfall in earnings that they will incur due to the price cap. Suppose that baseload generation is competitive:*
(i) With at most three states of nature (and hence at most two states of nature with market power), the Ramsey optimum can be achieved despite the presence of market power through a combination of price cap and capacity obligations, provided that

- *off-peak plants are eligible to satisfy LSE capacity obligations and to receive capacity payments,*
- *all consumers (including price-sensitive ones) are subject to the capacity obligations, and they pay the applicable capacity prices.*

(ii) With more than three states of nature (and more than two states of nature with market power), a combination of a price cap and capacity obligations is in general inconsistent with Ramsey optimality. The regulator faces a trade-off between alleviating market power off peak through a strict price cap and not overincentivizing peakers; and is further unable to provide price-sensitive consumers with proper economic signals in all states of nature.

Time inconsistency / political economy

(Coming back to perfect competition and two states of nature), suppose that the regulator imposes an unannounced *price cap*, $p^{\max} < p_2^*$, once K_2 has been sunk. A regulatory rule that sets a price cap equal to the marginal operating cost of the

[24]The consumers that regroup within a bubble must then design an internal market (with price p_2^*) in order to induce an internally efficient use of their global capacity obligations.

peaking unit with the highest marginal cost is an example. Such a rule precludes recovery of the scarcity rents needed to provide appropriate incentives for investment in peaking capacity. Then one would want a capacity payment to offset insufficient incentives:

$$p_K = f_2\left(p^* - p^{\max}\right).$$

The second best is then restored subject to the caveats enunciated in the previous subsection (except for the one on ex ante monopoly behavior, which is not relevant here).

The imposition of a price cap in this case is of course a hold-up on peak-load investments (peakers).[25] In practice, what potential investors in peaking capacity want is effectively a forward contract that commits to capacity payments to cover their investment costs to ensure that they are not held up ex post. They are comfortable that they have a good legal case that they can't be forced to produce if the price does not at least cover their variable production costs. It is the "scarcity rents" that they are concerned will be extracted by regulators or the ISO's market monitors.

Absence of clearing price

The third avenue is to assume a choke price: $\widehat{D}_2\left(p_2^*\right) = 0$ (the peak price goes up so much that no consumer under RTP ever wants to consume). Alternatively, one could consider the very, very short run, for which basically no-one can react (even the $\widehat{D}$ consumers). Either way, the supply and demand curves are both vertical and the price is infinite (given $D_2\left(p^*\right) > K_2$ under the first hypothesis).

One can set $p_2 =$ VOLL in order to provide generators with the right incentives in the absence of capacity payment. As Stoft (2002) argues, VOLL pricing augments

[25] Regulatory hold-ups may occur through other channels than price caps. For example, the ISO may purchase excessive peaking capacity and dispatch it at marginal operating cost during peak. We take up procurement issues in Section 3.2.

market power. But again, it is unclear whether market power is best addressed through price caps or through a requirement that LSEs enter into forward contracts for a large fraction of their peak demand or through some other mechanism. Another potential issue is that the regulatory commitment to VOLL pricing (that may reach 500 times the average energy price) may be weak. A third potential issue is that the VOLL is very hard to compute: As we discussed above, the outage cost for the consumer varies substantially with the degree of anticipation of the outage and its length.[26]

Whatever the reason, regulatory authorities most often set a price cap that lies way below (any reasonable measure of) the VOLL. As is well-known and was discussed earlier, the price cap depresses incentives for investment in peakers. Consumers and LSEs individually have no incentive to compensate for the peakers' shortfall in earnings to the extent that benefits from capacity investment are reaped by all (a free rider problem).

Thus, the analysis is qualitatively the same as previously; quantitatively, though, the effects are even more dramatic due to the very large wedge between the price cap and the socially optimal price during outages.

3.2 Procurement by the ISO

Another potential factor leading to discrepancies between wholesale prices and social scarcity values is linked to the way system operators purchase, dispatch and charge for energy and reserves (Patton 2002). We study the implications of two such practices: out-of-merit dispatching and recovery of some of the capital or operating costs of generation through an uplift charge that allocates these costs to wholesale prices based on some administrative costs allocation procedure.

As described by Patton, Van Schaik and Sinclair (2004, page 44) in their recent

[26] EdF (1994, 1995).

evaluation of the New England ISO's real time wholesale energy market,

> "Out-of-merit dispatching occurs in real time when energy is produced from a unit whose incremental energy bid is greater than the LMP [locational marginal price] at its location. In a very simple example, assume the two resources closest to the margin are a $60 per MWh resource and a $65 per MWh resource, with a market clearing price set at $65. When a $100 per MWh resource is dispatched out-of-merit, it will be treated by the [dispatch] software as a resource with a $0 [per MWh] offer. Assuming the energy produced by the $100 resource displaces all of the energy produced by the $65 resource, the [locational marginal] price will decrease to $60 per MWh."

Accordingly, the marginal cost of the most expensive resource dispatched is greater than the market clearing price and the associated marginal value placed on incremental supplies by consumers at its location. Note as well that in this example, the ISO effectively pays two prices for energy. It pays one price for energy dispatched through the market and a second higher price for the resource dispatched out-of-merit, while treating the latter in the dispatch stack as if it had a bid (marginal cost) of zero. Out-of-merit dispatch is typically rationalized as being necessary to deal with reliability constraints or dynamic factors related to minimum run-times or ramping constraints that are not fully reflected in the "products" and associated prices available to the ISO in its organized public markets.

Uplift refers to a situation in which the ISO makes a payment to a generator in excess of the revenues the generator would receive by making sales through the ISO's organized wholesale markets. These additional payments are then recovered by the ISO by placing a surcharge on wholesale energy transactions based on some administrative cost allocation formula. The costs of out-of-merit dispatch, the costs

of voltage support in the absence of a complete set of reactive power markets, out-of-market payments made by the ISO to ensure that specific generating units are available during peak demand periods, and out-of-market payments made by the ISO to certain customers to allow the ISO to curtail their demands on short notice may be recovered through uplift charges. In what follows, however, we treat the effects of out-of-merit dispatch and recovery through uplift charges separately. Different sources of uplift costs may be recovered with different allocation procedures (Patton, VanSchaick and Sinclair, page 51.)

3.2.1 Out-of-merit dispatching

In this subsection, we assume that the ISO contracts for peak production plants and dispatches them at the bottom of the merit order (at price 0), without regards to a price-cost test. Assume that there are two states: State 1 is off-peak, state 2 peak. K_1 is baseload capacity (investment cost I_1, marginal cost c_1), K_2 is peak capacity, used only during peak (investment cost $I_2 < I_1$, marginal cost $c_2 > c_1$). Consumers react to the real-time price. A fraction f_1 (resp. f_2) of periods is off peak, with demand $D_1(p)$ (resp. on peak, with demand $D_2(p) > D_1(p)$).

Competitive equilibrium (indexed by a "star"):

Free entry conditions:

$$I_1 = f_1 (p_1^* - c_1) + f_2 (p_2^* - c_1)$$

$$I_2 = f_2 (p_2^* - c_2)$$

Supply = demand:

$$D_1 (p_1^*) = K_1^*$$

$$D_2 (p_2^*) = K_1^* + K_2^* = K^*$$

The competitive equilibrium is depicted in figure 1.

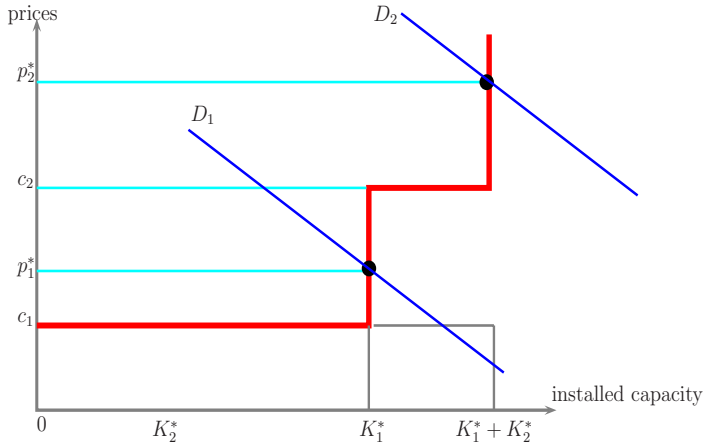

Figure 1

ISO procurement behavior

Suppose that the ISO contracts for $K_2^0 \leq K_2^*$ units of capacity and dispatches them at price 0 even off peak. This sounds strange, but more generally, as long as ISO purchases are financed externally, perverse effects arising from ISO dispatch decisions arise only if the dispatch is not economically efficient as long as $K_2^0 \leq K_2^*$. Note also that one could imagine that state 1 is an intermediate state of demand. There would then be an off-peak state 0 with frequency f_0. As long as the off-peak price p_0^* is unaffected, one can easily generalize the analysis below.

In order to clearly separate the effect studied here from that analyzed in the next subsection, assume that ISO losses (to be computed later) are financed externally (in practice, there would be injection / withdrawal taxes, that would shift the curves. Let us thus abstract from such complications).

Short-term impact. We analyze the short-term impact assuming a fixed capacity K_2^*. One may have in mind that K_2^0 of the K_2^* units of peaking capacity are purchased by the ISO. For given investments K_1^* and K_2^*, the short-term impact of the ISO

policy is depicted in figure 2, which assumes $K_2^0 = K_2^*$:

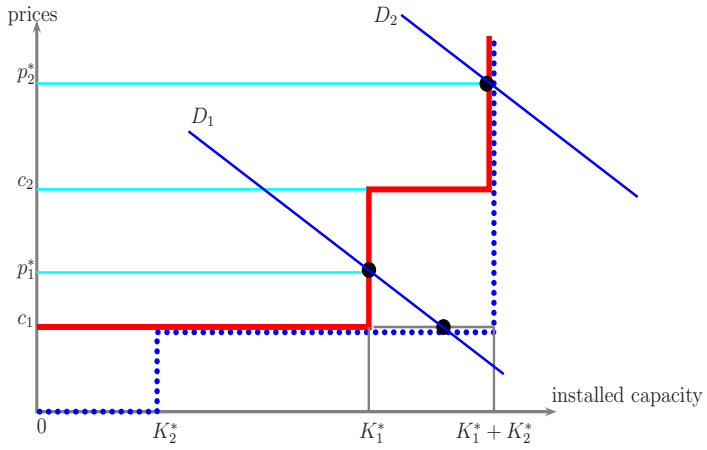

Figure 2

- the peak price remains unchanged (p_2^*),
- the off-peak price falls to $\max\left\{c_1, D_1^{-1}\left(K_1^* + K_2^*\right)\right\} = p_1^{ST}$,
- there is overproduction off-peak,
- the ISO loses

$$f_1 K_2^* \left(c_2 - p_1^{ST}\right).$$

Long-term effects. Suppose that the ISO buys a quantity $K_2^0 \leq K_2^*$ of peak-period units that it dispatches at zero price. It is easily seen that prices and capacities adjust in the following way:

- $p_2^{LT} = p_2^*$
- $p_1^{LT} = p_1^*$
- Peak units substitute partly for off-peak units (production inefficiency): $K_1^* - K_1^{LT} = K_2^0$ (or else $K_1^{LT} = 0$ if $K_2^0 \geq K_1^*$).

Proposition 4 *Suppose that ISO purchases $K_2^0 \leq K_2^*$ are financed externally (i.e., not through an uplift) and are dispatched out-of-merit.*
(i) The short-term incidence of a purchase is entirely on off-peak price and quantity: p_1 decreases, q_1 increases.
(ii) The long-term incidence of a purchase $K_2^0 \leq K_1^*$ is a substitution of off-peak units by peakers; on- and off-peak prices are unaffected.

Proof: To prove part (ii), note first that $p_2 > p_2^*$ is inconsistent with the free-entry condition. Next if $p_2 < p_2^*$, then $K = K_1 + K_2^0 > K^*$, and so if $K_2^0 \leq K_2^*$, $p_1 < p_1^*$; but then $K_1 = 0$, a contradiction. Hence $p_2 = p_2^*$. Next either $K_1 = 0$ or $K_1 > 0$. In the latter case, $p_1 = p_1^*$ by the free entry condition. To get this price, one must have $K_2^0 + K_1^{LT} = K_1^*$ (see figures 1 and 2). ■

Remark: The analysis in this section assumes that the ISO purchases no more than K_2^* units of peaking capacity and finances any revenue shortfalls externally. In this case inefficiencies come solely from inefficient dispatching. That is, there is no inefficiency as long as energy is dispatched only when the market price exceeds marginal cost. Moreover, peak period prices are unaffected even if dispatch is inefficient. However, if the ISO were to purchase more than K_2^* units of peaking capacity it could affect the peak period price even if the dispatch were efficient. Specifically if the ISO made additional purchases of peaking capacity to increase its ownership to more than K_2^* units and dispatched it efficiently only to meet peak period demand, the peak period price would fall in the short run. If it purchased a large enough quantity of additional peaking capacity and bid it into the market at its marginal cost c_2 it could drive the peak period price down to c_2. Clearly, such an ISO investment strategy would be inefficient. Moreover, such a strategy would have significant adverse long run effects on private investment incentives. Private investment in peaking capacity would be unprofitable and the incentives to invest in base load

capacity would also be reduced. In the long run this would lead to a substitution of peaking capacity for base load capacity and could potentially lead to a situation where the ISO had to purchase a large fraction of the capacity required to balance supply and demand.

3.2.2 Recovery through an uplift

In practice, ISO purchases are not financed through lump-sum taxation. Rather some or all of the associated costs are often at least partially recovered through an uplift. There is no general rule on how uplifts are recovered. They can be recovered monthly (often) or annually. They are typically spread across all kWh, but they can also be allocated to groups of hours (for example peak hours). In this section, we will work with the polar case assumption that none of the associated costs of ISO purchases are recovered from market revenues, but we recognize that some of these costs may be recovered from market revenues rather than uplift charges. There are several reasons for why some of the costs of ISO purchases in practice are not fully recovered from selling the energy in the market and so an uplift is needed: existence of a price cap; absence of a locational price allowing recovery at an expensive node; and usage of reserves outside the market place.

a) Let us analyze the implications of an uplift, starting with the *case in which the cost recovery is spread over peak and off-peak periods* (the cost is "socialized" through the uplift).

Suppose that the system operator purchases K_2^0 units of peaking energy forward, and dispatches the corresponding units only on peak (so that the inefficiency studied in subsection 3.2.1 does not arise). Total capacity to meet peak demand is then $K_1 + K_2$, where

$$K_2 = K_2^0 \quad \text{if} \quad f_2\left(p_2 - c_2\right) < I_2$$

$$K_2 \geq K_2^0 \quad \text{if} \quad f_2\left(p_2 - c_2\right) = I_2.$$

The uplift t is given by

$$t\left[f_1 D_1\left(p_1+t\right)+f_2 D_2\left(p_2+t\right)\right] \;=\; K_2^0 I_2$$

Off-peak capacity, K_2, and prices are given by:

$$D_1\left(p_1+t\right)=K_1$$

$$f_1\left(p_1-c_1\right)+f_2\left(p_2-c_1\right) \;=\; I_1$$

$$\Longrightarrow \;\; E\left(p\right)=E\left(p^*\right).$$

Peak capacity satisfies:

$$D_2\left(p_2+t\right) \;=\; K_1+K_2.$$

And so

$$t\left[K_1+f_2 K_2\right] \;=\; K_2^0 I_2.$$

Figure 3 depicts the equilibrium outcome for linear demands ($D_i\left(p\right)=a_i-p$). For small purchases K_2^0, production prices (p_1, p_2) don't move with the size of procurement. This is because the private sector still offers peaking capacity beyond K_2^0 and so peak and off-peak prices must remain consistent with the free-entry conditions. Investment in off-peak capacity is negatively affected by the uplift, while total peaking capacity does not move (the latter property hinges on the linearity of demand functions and is not robust). At some point, the private sector finds it uneconomical to invest in peakers; the only available peaking capacity is then that procured by the ISO. The peak price falls and the (before tax) off-peak price grows with the size of purchases.

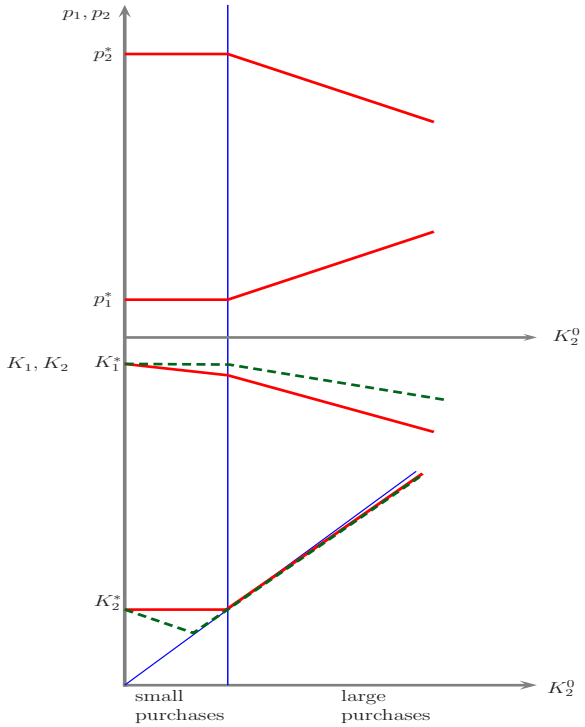

Figure 3: Socialized uplift (dotted line: uplift levied solely on peak consumption)

The results generalize to demand functions such that

$$D_2'(p_2) \leq D_1'(p_1) \text{ whenever } p_2 \geq p_1$$

(this condition is much stronger than needed, though).

b) Last, let us consider the impact of an *uplift levied solely in peak periods* .

The uplift, when levied on peak consumption only, is given by:

$$f_2 t D_2(p_2 + t) = K_2^0 I_2 \Longleftrightarrow f_2 t (K_1 + K_2) = K_2^0 I_2.$$

The *off-peak* conditions are

$$D_1(p_1) = K_1$$

and

$$f_1(p_1 - c_1) + f_2(p_2 - c_1) = I_1,$$

or, equivalently

$$E[p] = E[p^*].$$

The *peak* conditions are, as earlier:

$$K_2 = K_2^0 \quad \text{if} \quad f_2(p_2 - c_2) < I_2$$

$$K_2 \geq K_2^0 \quad \text{if} \quad f_2(p_2 - c_2) = I_2,$$

and

$$D_2(p_2 + t) = K_1 + K_2.$$

Hence:

$$D_2\left(p_2 + \frac{K_2^0 I_2}{f_2(K_1 + K_2)}\right) = K_1 + K_2. \tag{10}$$

We assume that the equation in K (for an arbitrary p_2)

$$D_2\left(p_2 + \frac{K_2^0 I_2}{f_2 K}\right) = K$$

admits a single solution K and that this solution is decreasing in K_2^0.[27]

For *small purchases*, as in the case of a socialized uplift, a small purchase K_2^0 is complemented by private sector offering ($K_2 > K_2^0$) and so $p_2 = p_2^*$. Given that the average price must be the same as for the free entry equilibrium, p_1 is then equal to p_1^*.

Hence, for K_2^0 small,

$$p_1 = p_1^* \quad \text{and} \quad p_2 = p_2^*$$

$$K_1 = K_1^*.$$

K_2 decreases as K_2^0 increases : There is *more than full crowding out of private investment in peakers by ISO purchases.*

For *larger purchases* at some point $K_2 = K_2^0$ and private investment in peakers disappears ($f_2(p_2 - c_2) \leq I_2$). But (10) still holds. Suppose that when K_2^0 increases, p_2 increases; then p_1 decreases (as the average price must remain constant) and so

[27] One has

$$\left[1 + D_2' \frac{K_2^0 I_2}{f_2 K^2}\right] dK = \frac{D_2' I_2}{f_2 K} dK_2^0.$$

Because, in this range, $I_2 = f_2(p_2 - c_2)$, a sufficient condition for this is that the peak elasticity of demand $-D_2' p_2 / D_2$ be equal to or less than one.

K_1 increases (and so does K). For a given K, the left-hand side of (10) decreases as p_2 and K_2^0 increase. So to restore equality in (10), K must decrease, a contradiction. Hence p_2 increases.

Proposition 5 *Suppose that an uplift is levied in order to finance ISO purchases, and that the latter are dispatched in merit.*
(i) If the uplift is socialized, off-peak capacity is reduced, peak capacity may increase or decrease, and prices are unaffected for small purchases. For larger purchases, the off-peak price increases while the off-peak capacity decreases; the peak price decreases while the peaking capacity increases with the size of the purchases. As ISO purchases increase, private investment in peakers becomes unprofitable at some point and the only available peaking capacity is procured by the ISO.
(ii) If the uplift applies solely to peak energy consumption, only peak capacity is affected (downward) for small purchases. For larger purchases, the characterization is the same as for a socialized uplift. There is more than full crowding out of peakers by ISO purchases and as ISO purchases increase a point is reached were private investment in peakers disappears.

4 Network support services and blackouts

This section relaxes another key assumption underlying our benchmark proposition (Proposition 1). There, we assumed that, while there may be insufficient resources and rationing, this rationing makes use of all available generation resources. This assumption is a decent approximation for, say, controlled rolling blackouts where the system operator sheds load sequentially to ensure that demand does not exceed available generating capacity. It is not for system collapses where deviations in network frequency or voltage lead to both generators and load tripping out by automatic protection equipment whose operation is triggered by physical disturbances on the network. For example, the August 14, 2003 blackout in the Eastern

United States and Ontario led to the loss of power to over 50 million consumers as the networks in New York, Ontario, Northern Ohio, Michigan and portions of other states collapsed. Over 60,000 MW of generating capacity was knocked out of service in a few minutes time. Most of the generating capacity under the control of the New York ISO tripped out despite the fact that there was a surplus of generating capacity to meet demand within the New York ISO's control area. Full restoration of service took up to 48 hours. (U.S.-Canada Power System Outage Task Force, 2003). The September 28, 2003 blackout in Italy led to a loss of power across the entire country and suddenly knocked out over 20,000 MW of generating capacity. Restoration of power supplies to consumers was completed about 20 hours after the blackout began (UCTE, 2003).

Conceptually, there is a key difference between rolling blackouts in which the system operator sequentially sheds relatively small fractions of total demand to match available supplies in a controlled fashion and a total system collapse in which both demand and generation shuts down over a large area in an uncontrolled fashion. Under a rolling blackout, available generation is extremely valuable (actually, its value is VOLL). By contrast, available plants are almost valueless when the system collapses. To put it differently, there is then an externality imposed by generating plants (or transmission lines) that initiate the collapse sequence on the other plants that trip out of service as the blackout cascades through the system, that does not exist in an orderly, rolling blackout.

It is useful here to relate this economic argument to standard engineering considerations concerning operating reserves (OpRes). In addition to dispatching generators to supply energy to match demand, system operators schedule additional generating capacity to provide operating reserves (OpRes). Operating reserves typically consist of "spinning reserves" which can be fully ramped up to supply a specified rate of electric energy production in less than 10 minutes and "non-spinning reserves"

which can be fully ramped up to supply energy in up to 30 minutes (60 minutes in some places). Operating reserves are used to respond to sudden outages of generating plants or transmission lines that are providing supplies of energy to meet demand in real time sufficiently quickly to maintain the frequency, voltage and stability parameters of the network within acceptable ranges. Additional generation is also scheduled to provide continuous frequency regulation (or automatic generation control) to stabilize network frequency in response to small instantaneous variations in demand and generation. These ancillary network support services require scheduling additional generating capacity equal to roughly 10-12% of electricity demand at any point in time. In the U.S., regional reliability councils specify the requirements for frequency regulation and operating reserves, as well as other ancillary services such as reactive power supplies and blackstart capabilities, that system operators are expected to maintain. Pending U.S. legislation would make these and other reliability standards mandatory for system operators.

Let us use a simple model of OpRes in order to analyze the various issues at stake. To keep modeling details to a minimum without altering insights, the demand side is modeled as inelastic: In state $i \in [0,1]$, demand is D_i. If $d_i \leq D_i$ is served, the consumers' gross surplus is $d_i v$, where v is the value per kWh (the value of lost load). Similarly, on the supply side, there is a single technology: capacity K involves investment cost KI and marginal cost c, which we normalize at 0 in order to simplify accounting.

The key innovation relative to the benchmark model is that the extent of scarcity is not fully known at the time that generating units with uncertain availability are scheduled to meet the demand dispatched by the system operator. We formalize this uncertainty as an uncertain capacity *availability* factor $\lambda \in [0,1]$. That is, a fraction $1-\lambda$ of the capacity K will break down. The distribution $H_i(\lambda)$(with $H_i(0) = 0$ and

$H_i(1) = 1$) can be state-contingent.[28] There may be an atom in the distribution at $\lambda = 1$ (full availability), but the distribution has otherwise a smooth density $h_i(\lambda)$.[29] We make the following weak assumption:

$$\frac{h_i(\lambda)\,\lambda}{[1 - H_i(\lambda)]} \quad \text{is increasing in } \lambda$$

(a sufficient condition for this is the standard assumption that the hazard rate $h_i/[1 - H_i]$ is increasing in λ, which is satisfied for most commonly used continuous probability distributions.)

The timing goes as in Figure 4: Given nominal capacity K and demand D_i, the SSO chooses how much of this demand to dispatch , or alternatively how much demand to curtail, and a reserve margin. More formally, once load D_i is realized, the system operator can curtail an amount $D_i - d_i \geq 0$ of load. He also chooses a reserve coefficient r_i, so that a capacity $(1 + r_i)\,d_i \leq K$ must be ready to be dispatched. Then, the capacity availability λ_i is revealed and the demand $d_i \leq D_i$ is served or the network collapses: If $\lambda_i\,[(1 + r_i)\,d_i] < d_i$, the system collapses, and no energy is produced or consumed.

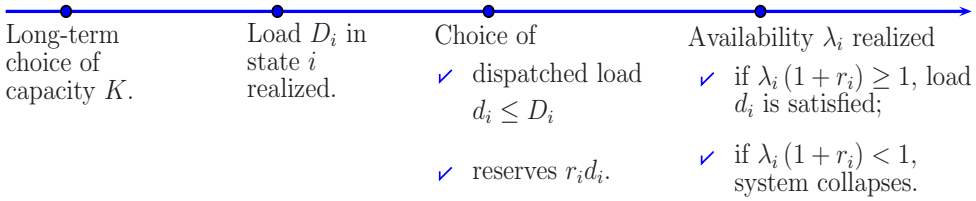

Figure 4

We assume that scheduling generation to be (potentially) available to serve demand costs s per unit (s can be either a monetary cost of keeping the plant ready

[28] For example, if plant unavailability comes from the breakdown of a transmission line connecting the plant and the load, the transmission line may be more likely to break down under extreme weather conditions, for which load D_i is also large.

[29] We assume a continuous distribution solely for tractability purposes. In practice, system operators fear foremost the breakdown of large plants or transmission lines and therefore adopt reliability criteria of the type "$n - 1$" or "$n - 2$". This introduces "integer problems", but no fundamental difference in analysis.

to be dispatched or an opportunity cost of not being able to perform maintenance at an appropriate time).

a) *Social optimum*

A Ramsey social planner would solve:

$$\max_{\{K,d.,r.\}} \left\{ E\left[\left[1-H_i\left(\frac{1}{1+r_i}\right)\right]v - s\left(1+r_i\right)\right]d_i - KI \right\}$$

such that, for all states $i \in [0,1]$:

$$d_i \leq D_i \qquad (\mu_i)$$

$$\left(1+r_i\right)d_i \leq K, \qquad (\nu_i)$$

where μ_i and ν_i are the shadow prices of the constraints.

For conciseness, we analyze only the case where it is optimal to accumulate reserves in each state. The first-order conditions with respect to r_i, d_i and K are, respectively:

$$\frac{h_i}{\left(1+r_i\right)^2}v - s = \nu_i, \tag{11}$$

$$\left[1-H_i\right]v - s\left(1+r_i\right) \geq \mu_i + \left(1+r_i\right)\nu_i, \quad \text{with equality unless } d_i = D_i \tag{12}$$

and

$$E\left[\nu_i\right] = I. \tag{13}$$

Specializing the model to the case in which H_i is state-independent,[30] let us analyze the optimal dispatching, as described by (11) and (12). The Ramsey optimum is depicted in Figure 5.

[30] We will still use state-denoting subscripts, though, so as to indicate the value taken for H in state i. For example, $H_i = H\left(1/\left(1+r_i\right)\right)$.

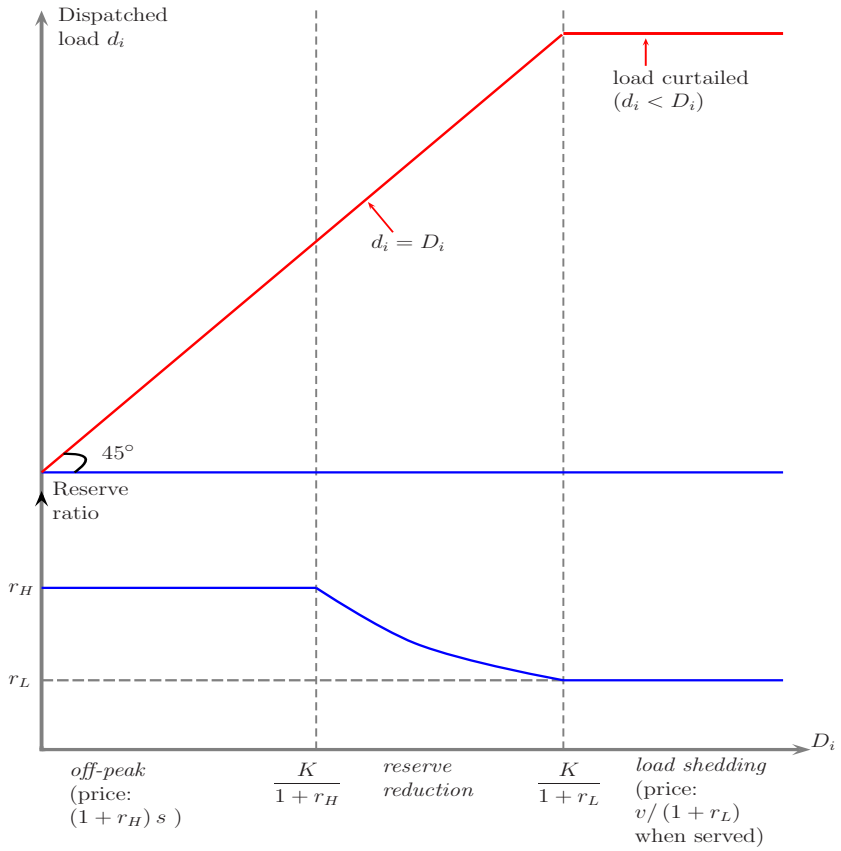

Figure 5 (prices indicated in parentheses are prices paid by consumers)

Off-peak (D_i small), there is excess capacity, all demand is served ($d_i = D_i$), and $\nu_i = 0$. Hence from (11) and (12)

$$r = r_H$$

where

$$\frac{h\left(\frac{1}{1+r_H}\right)}{(1+r_H)^2} v = s.$$

We of course assume that for this value, it is worth dispatching load ($\mu_i > 0$), or

$$\left[1 - H\left(\frac{1}{1+r_H}\right)\right] v > s\,(1+r_H)\,.$$

The off-peak region is defined by:

$$(1+r_H)\, D_i < K.$$

Peaking time can be decomposed into two regions. As D_i grows, load first keep being satisfied: $d_i = D_i$, and reserves become leaner (and so the probability of a

blackout increases as load grows):

$$(1 + r_i) D_i = K.$$

Load starts being shed when $\mu_i = 0$, or

$$\frac{h_i}{[1 - H_i]} \cdot \frac{1}{1 + r_i} = 1,$$

which from our assumptions has a unique solution:

$$r_L < r_H.$$

The optimal investment policy is then given by:

$$I = \int_{\frac{K}{1+r_H}}^{\frac{K}{1+r_L}} \left[\frac{h_i}{(1 + r_i)^2} v - s \right] f_i di + \int_{\frac{K}{1+r_L}}^{\infty} \left[(1 - H_i) \frac{v}{(1 + r_L)} - s \right] f_i di.$$

The first term on the right-hand side of this equation represents the quasi-rents in reserve reduction states: An extra unit of capacity is used to increase reserves and thereby reduce the probability of network collapse, $1 - H_i (D_i/K)$, then saving value of lost load vD_i; to this term must be subtracted the cost s of scheduling generation. And the second term represents the quasi-rents in load shedding states: An extra unit of capacity allows a reduction in load shedding, which has value $(1 - H_i) v/(1 + r_L)$ minus the cost s of scheduling generation.

b) *Implementation*

First, note that the possibility of system collapses make operating reserves a public good. Network users take its reliability as exogenous to their own policy and thus are unwilling to voluntarily contribute to reserves. The market-determined level of reliability is therefore the size of the atom of the $H(\cdot)$ distribution at $\lambda = 1$. Thus, the market solution leads to an insufficient level of reliability. In order to obtain a proper level of reliability, the system operator must force consumers (or their LSE) to purchase a fraction r_i of reserves for each unit of load.[31]

[31] There is no point further asking generators to hold reserves.

Does this market mechanism cum regulation of reserve ratios generate enough quasi-rents to induce the optimal investment policy? Off-peak ($D_i < K/(1+r_H)$), the price paid by consumers for reserves is $(1+r_H)\,s$, and there are no quasi-rents.

When load is curtailed ($D_i > K/(1+r_L)$), then consumers must pay $\mathrm{v}/(1+r_L)$ conditionally on being actually served (which has probability $1-H_i$). Thus, generators obtain, as they should, quasi-rent:

$$(1-H_i)\,\frac{\mathrm{v}}{1+r_L} \;-\; s$$

in this region.

The intermediate region is more complex to implement through an auction-type mechanism. In the absence of price-responsive load, the supply curve and the total demand curve (energy plus reserves) are vertical and identical. Hence a small mistake in the choice of reserve ratio creates wild swings in the market price (from $(1+r_i)\,s$ to $\mathrm{v}/(1+r_i)$ conditionally on being served). In particular, the system operator can bring price down to marginal cost without hardly affecting reliability. This has potentially significant implications for investment incentives.

The "knife edge" problem has been recognized by system operators. It puts a lot of discretion in the hands of the system operator to affect prices and investment incentives as small deviations in this range can have very big effects on prices. In the end, determining when there is an operating reserve deficiency (or a forecast operating reserve deficiency) may necessarily involve some discretion because it depends in part on attributes of the network topology that are not reflected in a refined way in the rough requirements for operating reserves (e.g. ramp up in less than 10 minutes). So, for example, stored hydro is generally thought to be a superior source of operating reserves than fossil plants because the former can be ramped up almost instantly rather than in 9 minutes. If there is a lot of hydro in the OpRes portfolio the system operator will be less likely to be concerned about a small shortfall in

operating reserves.

Alternatively, the system operator can compute the marginal social benefit, $\left(h'\frac{D_i}{K^2}\right) \cdot (D_i \mathrm{v})$, of the reduction in the probability of collapse brought about by an additional unit of investment. This regulated price for reserves (and thus for energy) then yields the appropriate quasi-rent:

$$\frac{h_i}{(1+r_i)^2}\,\mathrm{v} - s$$

to generators in this region. Accurately computing the regulated prices in this region also involves substantial discretion, however.

Proposition 6 *Suppose that the extent of scarcity is not known with certainty at the time of generator and load dispatch.*
(i) The socially optimal policy involves, as the forecasted demand grows, three regimes:

- *Off peak: the entire load is dispatched, and operating reserves are set at a fixed, maximum percentage of load.*
- *Reserve shedding: the entire load is dispatched, and operating reserves are reduced as generation capacity is binding.*
- *Load shedding: Load is curtailed, and operating reserves satisfy a fixed, minimum ratio relative to load.*

(ii) The possibility of system collapses makes operating reserves a public good. As a result, investments in operating reserves do not emerge spontaneously as a market outcome. The load should be forced to pay for a pre-determined quantity of operating reserves (e.g. as a proportion of their demand) :

- *a price set at VOLL (divided by one plus the reserve ratio, conditionally on being served) in the load shedding region,*
- *a market clearing price given the ratio requirement off peak,*

- *a price growing from marginal cost to the load-shedding-region price in the reserve-shedding region. Decentralization through an operating reserves market together with a mandatory reserve ratio is delicate as the price of reserves is extremely sensitive to small mistakes or discretionary actions by the system operator.*

5 Conclusion

We derived the (second-best) optimal program for prices, output and investment for an electricity sector in which non-price sensitive consumers may have to be rationed under some contingencies. This allocation provides a benchmark against which the actual performance of electricity sectors, and the effects of the imposition of various regulatory and non-market mechanisms and constraints, can be compared. We went on to show that competitive wholesale and retail markets will support this second-best "Ramsey" allocation under a particular set of assumptions.

The assumptions underpinning these results are very strong. Our research program seeks to evaluate the effects of departures from the assumptions needed to support the benchmark allocation. In this paper we focused on relaxing the assumptions (a) that wholesale electricity prices reflect the social opportunity cost of generation and (b) that rationing, if any, is orderly and makes efficient use of available generation.

To examine the effects of relaxing the first assumption, we analyzed the effects of regulator-imposed prices caps motivated either by concerns about market power in the real time market or by regulatory opportunism. While price caps can significantly reduce the scarcity rents required to cover the costs of investment in peaking capacity, lead to underinvestment, and distort the prices seen by consumers, with at most three states of nature (and up to two states with market power), capacity obligations and associated capacity payments can restore investment incentives if

all generating capacity is eligible to meet capacity obligations and receive capacity payments, and all consumer demand is subject to capacity obligations. We go on to examine the effects of ISO procurement strategies that involve either inefficient generator dispatch or the recovery of some generation costs through an uplift. These ISO procurement strategies can distort prices and investment decisions in various ways.

Our analysis then proceeded to examine the effects of relaxation of the second assumption underpinning the benchmark allocation. We used a model of uncertain demand and operating reserves to analyse the effects of network collapses that result in rationing of demand while generation that is potentially available to meet this demand stands idle. Unlike the benchmark model, the extent of scarcity is not known with certainty at the time of generator dispatch. In this model operating reserves are a public good and without mandatory operating reserve requirements there would be under investment in operating reserves and lower reliability than is optimal. Moreover, under certain contingencies the market price, and the associated scarcity rents available to support investments in generating capacity, are extremely sensitive to small mistakes or discretionary actions by the system operator. This "knife edge" problem and options for dealing with it requires further analysis and attention in the development of the rules and incentive arrangements governing system operators.

In Joskow-Tirole (2004) we examine relaxation of the other key assumptions that underpin the benchmark model, focusing on the impacts of load profiling, zonal rationing of demand for both price sensitive and price insensitive consumers, and more general characterizations of consumer heterogeneity. Taken together, these results suggest that the combination of the unusual physical attributes of electricity and electric power networks and associated reliability considerations, limitations on metering of real time consumer demand and responsiveness to real time prices,

restrictions on the ability to ration individual consumers, discretionary behavior by system operators, makes achieving an efficient allocation of resources with competitive wholesale and retail market mechanisms a very challenging task.

References

[1] Allaz, B.(1992) "Uncertainty and Strategic Forward Transactions," *International Journal of Industrial Organization*, 10: 297–308.

[2] Allaz, B. and J.L. Vila (1993) "Cournot Competition, Forward Markets and Efficiency," *Journal of Economic Theory*, 59(1):1–16.

[3] Borenstein, S., and S. Holland (2003) "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices," CSEM WP116.

[4] Borenstein, S., Jaske, M., and A. Rosenfeld (2002) "Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets," Hewlett Foundation Energy Series.

[5] Chao, H. P., and R. Wilson (1987) "Priority Service: Pricing, Investment, and Market Organization," *American Economic Review*, 77: 899–916.

[6] —— (2003) "Resource Adequacy and Market Power Mitigation via Option Contracts," mimeo, Stanford University.

[7] EdF (1994) "The Explicit Cost of Failure," mimeo, General Economic Studies Department.

[8] —— (1995) "A New Value for the Cost of Failure,"mimeo, General Economic Studies Department.

[9] Green, R. (1999) "The Electricity Contract Market in England and Wales," *Journal of Industrial Economics*, Vol 47(1): 107–124.

[10] Joskow, P. and J. Tirole (2004) "Retail Electricity Competition," mimeo, MIT and IDEI.

[11] Littlechild, S. (2000) "Why We Need Electricity Retailers: A Reply to Joskow on Wholesale Spot Price Pass-Through," mimeo.

[12] Newbery, D. (1998) "Competition, Contracts and Entry in the Electricity Spot Market," *Rand Journal of Economics*, 29(4): 726–49.

[13] Oren, S. (2003) "Ensuring Generation Adequacy in Competitive Electricity Markets," mimeo, UC Berkeley.

[14] Patton, D. (2002) "Summer 2002 Review of the New York Electricity Market," presentation to the New York ISO Board of Directors and Management Committee (October 15).

[15] Patton, D., VanSchaick, P. L., and R. A. Sinclair (2004) "Six Month Review of the SMD Markets in New England," *Potomac Economics for the New England ISO*, February.

[16] Stoft, S. (2002) *Power System Economics*, Wiley.

[17] —— (2003) "The Demand for Operating Reserves: Key to Price Spikes and Investment," *IEEE*.

[18] Turvey, R.(2003) "Profiling: A New Suggestion," mimeo.

[19] Turvey, R., and D. Anderson (1977) *Electricity Economics: Essays ands Case Studies*. A World Bank Research Publication, Johns Hopkins University Press (Baltimore and London).

[20] Union for the Coordination of Electricity Transmission (UCTE) (2003) "Interim Report of the Investigation Committee of the 28 September 2003 Blackout in Italy," October 27.

[21] U.S.- Canada Power System Outage Task Force (2003) "Interim Report: Causes of the August 14 Blackout in the United States and Canada," November.

[22] Vasquez, C., Rivier, M., and I. Perez-Arriaga (2001) "A Market Approach to Long-Term Security of Supply," mimeo ITT, Universidad Pontificia Comillas, Madrid.

[23] Wolak, F. (2000) "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market," *International Economic Journal*, 14(2): 1–39.