Reliability and competitive electricity markets*

Paul Joskow † Jean Tirole‡

June 20, 2006

Abstract

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 (second-best) optimal prices and investment program when there are price-insensitive retail consumers, but when 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 examines the implications of relaxing several of these assumptions. First, it analyzes the interrelationships between regulator-imposed price caps and capacity obligations. 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.

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 on energy supplies, generating capacity contracting obligations placed on electricity distribution companies and other load serving entities (LSEs)[1], and system operating reserve requirements.

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, by non-price rationing of demand, by wholesale market power problems and by imperfections in mechanisms adopted to mitigate these market power problems.

*We are grateful to Hung Pao Chao, Claude Crampes, Ray Deneckere, Richard Green, Stephen Holland, Bruno Jullien, Alvin Klevorick, Patrick Rey, an anonymous referee, and the participants at the IDEI-CEPR conference on "Competition and Coordination in the Electricity Industry," January 16–17, 2004, Toulouse, the ninth annual POWER conference, UC Berkeley, March 19, 2004, and the EdF-R&D February 1, 2005 workshop in Paris for helpful discussions and comments. Alex Kendall and Patrick Warren provided able research assistance.

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

‡IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, and MIT.

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

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 requirements 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 (2005) is to start to bridge this gap.

The institutional environment in which our analysis proceeds has competing LSEs that market electricity to residential, commercial and industrial ("retail") consumers. Some retail consumers served by competing LSEs see and can respond to real time wholesale energy 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.[2] Retail consumers may be subject to non-price rationing to balance supply and demand in real time. The wholesale market is composed of generators who compete to sell power to LSEs. LSEs in turn compete to resell this power to retail consumers using the transmission and distribution delivery infrastructure. The prices for these delivery services are regulated and have been unbundled from power supply services. In what follows we normalize the prices for delivery services to zero. 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.

Section 2 first derives the optimal prices and investment program for an electric power system when there is state contingent demand, at least some consumers do not react to real time prices,

[2]Retail consumers may also be on a real-time meter, and yet only partly responsive to the real-time price due to transaction costs of adjusting consumption. As shown in Joskow-Tirole (2005), such consumers for the purpose of positive and normative analyses, can be treated as reactive consumers as long as they rationally trade off transaction costs and savings in their electricity bill. In particular, Proposition 1 still applies.

but their LSE can choose any level of rationing it prefers contingent on real time prices. The latter assumption is important. It implies that even if retail consumers do not have real time meters and cannot see or react directly to the real time price, LSEs which do see real time prices can enter into price-contingent priority rationing contracts (e.g. interruptible contracts) with the retail consumers with whom they have power supply contracts (as in Chao-Wilson 1987). We consider the effect of relaxing the assumption that individual retail consumers' consumption can be physically controlled by the system operator later in the paper and discuss the implications on zonal rationing in the presence of retail competition in Proposition 3 below and in Joskow-Tirole 2005. 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).[3] 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. That is, LSEs can enter into price contingent rationing contracts with their retail customers (alternatively we will allow the system operator to interrupt consumers to the best of their (ex ante) interests).
(e) Consumers have the same load profile[4] *(they are identical up to a scale factor).*

The assumptions underlying this benchmark proposition are obviously very strong: (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

[3] Joskow-Tirole (2005) analyzes more complex characterizations of consumer heterogeneity in the presence of retail competition.

[4] The phrase "load profile" here refers to the individual consumer's level of demand in each hour of the year relative to her peak hour demand.

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;[5] (d) price-sensitive consumers may be rationed along with everyone else who 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 (2005), 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 investment in and production from peaking capacity that runs during peak demand periods, and a price cap is applied that constrains the wholesale market price of energy to be lower than the competitive price during peak periods. This creates a shortage of peaking capacity in the long run. We show that placing generating capacity obligations on LSEs, combined with the associated capacity prices paid to generators 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 on energy produced from this capacity for the wholesale market. Indeed, with up to two states of nature with market power, the Ramsey optimum can be achieved despite the presence of market power through a combination of a price cap on energy and capacity obligations and associated capacity prices 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 two states of nature with market power, a combination of spot wholesale market price caps and capacity obligations will not achieve the Ramsey optimum. 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 investment incentives to meet peak demand efficiently, 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 on energy and a capacity price.

Section 4 derives the implications of network collapses and the concomitant need for net-

[5] By "load profiled" we mean that each consumer is billed based on the average load profile for a sample of consumers with similar attributes whose consumption is metered on a real time basis.

work support services that are typically provided by generating plants that the system operator schedules as "operating reserves". 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.[6] Hence, system collapses, unlike, say, controlled 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 with price-sensitive and price-insensitive consumers

2.1 Model

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.[7] We assume that the (unrationed) demand functions of price-insensitive and price-sensitive consumers, D_i and $\widehat{D}_i$, are increasing in i.[8]

Price-insensitive consumers are on traditional meters that record only their aggregate consumption over all states of nature, and therefore they do not react to the real time price (RTP).[9] Consumers are homogeneous, up to possibly a scaling factor, i.e., they have the same load profile.[10] 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 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, as,

[6] An analogy may help in understanding the distinction between orderly rationing and a collapse: 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 mad cow disease may spoil the entire market for beef.

[6] 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.

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

[8] 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.

[9] As in Joskow-Tirole (2005), we could also introduce consumers on real-time meters who do not monitor the real-time price. This would not affect Proposition 1 below.

[10] 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.

for example, when the system operator implements rolling blackouts. 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;[11] a case in point is that of consumers with interrruptible contracts. 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 $\mathcal{S}_i(D_i(p), \alpha_i)$.[12] More generally however, the consumer may adjust her demand to the prospect of being potentially rationed.[13]

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 conditions change relative to the reactivity of consumers.[14]

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

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

[11] 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.

[12] The surplus is linear in α_i in the perfectly anticipated blackout case discussed immediately below. However, more general expressions are feasible. For example, it is affine in α_i in case (a) of the opportunity cost example discussed subsequently.

[13] 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.

[14] This observation is made for example in Electricité de France (1994, 1995).

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 and the consumer gets stuck in the elevator. 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 a piece of 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) 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 $S_i^n(p) \equiv \max_D \{S_i(D) - pD\}$ 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

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

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

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

So, with perfectly anticipated blackouts, which give consumers time to adapt their behavior in anticipation of being curtailed, the value of lost load is equal to the average gross consumer surplus. It is higher for unanticipated blackouts than for planned blackouts.

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 for price-sensitive consumers; 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 $\widehat{\mathcal{D}}_i(\widehat{p}_i, \widehat{\alpha}_i)$ and their gross surplus is $\widehat{\mathcal{S}}_i(\widehat{p}_i, \widehat{\alpha}_i)$, 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.[15] There are constant returns to scale for each technology. We denote by $G(c) \geq 0$ the cumulative distribution function of plants.[16] So, the total investment cost is

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

The ex post cost of producing Q_i 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.

[15] For the undominated technologies, $I(c)$ is decreasing in c.

[16] 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:

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

s.t.

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

Letting $p_i f_i$ 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*:[17]

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

c) *Price-insensitive consumers*:

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

$$\text{(ii)} \qquad \text{Either} \quad \frac{\frac{\partial \mathcal{S}_i}{\partial \alpha_i}}{\frac{\partial \mathcal{D}_i}{\partial \alpha_i}}=p_i \quad \text{or} \quad \alpha_i\in\{0,1\}. \tag{3}$$

d) *Investment*:

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

These first-order conditions can be interpreted in the following way: condition (1) says that only those 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

[17] To prove condition (2), apply first the observation that by definition $\mathcal{D}_i(p_i,\widehat{\alpha}_i)$ 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:

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

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

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 ($\alpha_i \equiv 1$):

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

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

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

For example, for *perfectly foreseen outages*, using the consumer's first-order condition for the choice of $D_i(p)$, $\partial \mathcal{S}_i / \partial D_i = \alpha_i p$, condition (3) boils down to:

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

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

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

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 price-insensitive consumers are served by competitive LSEs, that LSEs face the real time wholesale price for the aggregate consumption of the retail customers for whom they are responsible, and that they can demand any level of state-contingent rationing $\alpha_i(p_i)$ for their consumers. This latter assumption implies that LSEs can enter into priority rationing contracts with their consumers (as in Chao-Wilson 1987).

The following proposition shows that, despite 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:

(a) the RTP reflects the social opportunity cost of generation,

(b) available generation is made use of during rationing periods,

(c) load-serving entities face the RTP,

(d) 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 α_i,[18] by, for example, auctioning price-contingent interruptible contracts to retail consumers.
(e) consumers have homogeneous demand and surplus functions (possibly up to a scaling factor).

Proof: Suppose that competing retailers (LSEs) can offer to price-insensitive contracts $\{A, p, \alpha_\cdot\}$, 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:

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

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. Q.E.D.

Remark: Chao-Wilson (1987) also emphasize the use of bids for priority servicing. 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 may involve 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) $\widehat{S}_1(\widehat{D}_1(p))$ and $\widehat{S}_2(\widehat{D}_2(p))$. We assume that rationing may occur only at peak ($\alpha_1 = 1\ ,\ \alpha_2 \leq 1$). In this two-state example p^* denotes the optimal non-contingent (constant) marginal price faced by price-insensitive retail consumers and p_1^* and p_2^* denote the optimal off-peak and peak (contintent) marginal prices, respectively.

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 $I_2 < I_1$. The marginal operating cost of the peakers is $c_2 > c_1$.

Social optimum: Letting p^* denote the (constant) price faced by price-insensitive consumers,

[18] 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.

the (second-best) social optimal solves over $\left\{p^*, \alpha_2, \widehat{D}_1, \widehat{D}_2\right\}$

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

where

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

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

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

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

$$f_1(p^* - p_1) D'_1 + f_2\left(\tfrac{\partial \mathcal{S}_2}{\partial p} - p_2 \tfrac{\partial \mathcal{D}_2}{\partial p}\right) = 0 \tag{3'i}$$

and

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

$$f_2(p_2 - c_2) = I_2 \Longleftrightarrow p_2 = c_2 + (I_2/f_2) \quad \text{(as long as } K_2 > 0\text{).}$$

Let $p_1^* < p^* < p_2^*$ be defined by:

$$p_2^* \equiv c_2 + (I_2/f_2)$$

and

$$p^* \equiv f_1 p_1^* + f_2 p_2^* \equiv I_1 + c_1.$$

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

To illustrate these conditions, let us focus on the case of:

Perfectly foreseen rolling blackouts: $\mathcal{S}_2(p, \alpha_2) = \alpha_2 S_2(D_2(p))$.

In this case, for an available peak capacity $K_1 + K_2$ and a pre-set retail price p charged to price-insensitive consumers, the optimal allocation of capacity among the price-sensitive and -insensitive consumers is given by:

$$\begin{aligned}&\max_{\{p_2,\, \alpha_2\}} \left\{\widehat{S}_2\left(\widehat{D}_2(p_2)\right) + \alpha_2 S_2(D_2(p))\right\} \\ &\text{s.t.} \\ &\widehat{D}_2(p_2) + \alpha_2 D_2(p) \leq K_1 + K_2.\end{aligned}$$

Let $\overline{p}_2(p)$ denote the solution to:

$$S_2\left(D_2(p)\right) = p_2 D_2(p).$$

Then:

$$\begin{array}{rcl} \alpha_2 & = & 0 \quad \text{if} \quad p_2 > \overline{p}_2(p) \\ & = & 1 \quad \text{if} \quad p_2 < \overline{p}_2(p) \\ & \in & [0,1] \quad \text{if} \quad p_2 = \overline{p}_2(p). \end{array}$$

The maximum spot price at which price-insensitive consumers ought to be served, $\overline{p}_2(p)$, is an increasing fucntion of p: A higher retail price reduces demand and thereby makes average consumption more desirable.

We can now endogenize the retail price:

(a) *Interruptability regime* $(\alpha_2 = 0)$:

In this regime the price charged to price-insensitive consumers is equal to the off-peak price $(p = p_1^*)$ and the retail consumers' surplus is $f_1\left[S_1\left(D_1(p_1^*)\right) - p_1^* D_1(p_1^*)\right]$. This requires that:

$$p_2^* > \overline{p}_2(p_1^*).$$

(b) *No-interruptability regime* $(\alpha_2 = 1)$:

The optimal price charged to price-insensitive consumers maximizes expected surplus:

$$\begin{array}{ll} \max\limits_{\{p\}} & \left\{f_1\left[S_1\left(D_1(p)\right) - p_1^* D_1(p)\right] + f_2\left[S_2\left(D_2(p)\right) - p_2^* D_2(p)\right]\right\} \end{array}$$

yielding:

$$E\left[\left(p - p_i^*\right) D_i'\left(p\right)\right] = 0. \qquad (6')$$

This expression takes a particularly simple form for *linear demands*:

$$D_i(p) = a_i - bp.$$

Then

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

This regime requires that:

$$p_2^* < \overline{p}_2\left(p^*\right).$$

Let us for example vary the peak cost parameters c_2 or I_2, and thereby vary p_2^* below $\overline{p}_2\left(p^*\right)$ (the latter being exogenously given by $p^* = I_1 + c_1$). As p_2^* decreases, p_1^* increases. And so there is indeed a level at which $p_2^* = \overline{p}_2\left(p_1^*\right)$. It is easy to show that there exists $\widehat{p}_2$ such that:

- for $p_2^* < \widehat{p}_2$, non-interruptability is optimal,
- for $p_2^* > \widehat{p}_2$, interruptability is optimal.

Furthermore

$$\overline{p}_2\left(\frac{I_1 + c_1 - f_2 p_2^*}{f_1}\right) < \widehat{p}_2 < \overline{p}_2(p^*).^{19}$$

We can therefore state:

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

The same analysis offers some insight as to the impact of retail competition when there is only zonal rationing. Let us assume that LSEs are charged for the real-time consumption of their customers, but that consumers cannot be interrupted individually (so the fourth assumption in Proposition 1 is violated). That is, the physical topology of the distribution network only allows the system operator physically to ration groups of consumers connected to the same distribution feeder — what we refer to in Joskow-Tirole 2005 as "zonal rationing." Instead, the system operator optimizes over the interruptability decision α_2 in a given geographical zone as a reaction to the "average retail contract" offered by the LSEs. Whenever

$$\overline{p}_2(p_1^*) < p_2^* < \overline{p}_2(p^*),$$

there are two equilibria in retail competition: If all LSEs offer marginal price p_1^*, then the system operator will interrupt price-insensitive retail customers, whose consumption is expensive and not that valuable in per unit terms. Anticipating interruptability, LSEs offer $p = p_1^*$. Similarly, $p = p^*$ is also an equilibrium. This multiplicity of equilibria, together with the earlier characterization of the optimal regime, shows that the marginal price may be too low or too high under retail competition.[20]

Proposition 3 Suppose that consumers can be curtailed only at the zonal, rather than individual, level and that the system operator's curtailement decisions maximize consumer welfare. Then LSE competition results in two equilibria, one at price p_1^* and the other at price p^*, provided that

$$\overline{p}_2(p_1^*) < p_2^* < \overline{p}_2(p^*).$$

[19] To see this, note that the maximand in the non-interruptability regime decreases with p_2^*. At $p_2^* = \overline{p}_2(p^*)$, the maximand is equal to $f_1[S_1(D_1(p^*)) - p_1^* D_1(p^*)] < f_1[S_1(D_1(p_1^*)) - p_1^* D_1(p_1^*)]$. And conversely for the interruptability regime.

[20] This multiple equilibrium problem would not emerge if the system operator committed to a rationing policy ex ante (when retail contracts are designed). Joskow-Tirole 2005 examine various aspects of zonaI rationing in this context.

3 Price caps on energy, capacity obligations, and capacity prices

As discussed in the introduction, the assumptions underlying Proposition 1 are very strong. We now proceed to examine the consequences of three attributes of electricity markets in practice: (a) market power on the one hand, and price caps and other policy interventions on the other hand that create departures of real time prices from the social opportunity cost of generation; (b) available generation does not supply load or receive payments during blackouts associated with a network collapse; (c) technological constraints in the distribution network imply that consumers who are sensitive to real time prices may be blacked out along with everyone else and, relatedly, LSEs cannot generally demand any level of rationing they desire.

3.1 Capacity obligations and capacity prices

The attributes of an electric power system's power supplies are often characterized by two primary variables. The first is the installed generating capacity on the system. Generating capacity measures the instantaneous physical capability of generating facilities to produce electrical energy. The quantity of generating capacity can be increased with additional investment and declines when generating plants are retired.[21] The second variable is the quantity of electrical energy supplied at each point in time by this generating capacity. It is electrical energy that is of value to consumers. Since electricity cannot be stored the quantity of generating capacity must be at least as large as peak demand to avoid blackouts. And since demand varies widely from hour to hour during the year, the quantity of electricity produced relative to the stock of generating capacity on the system varies widely over the hours of the year as well.

Spot wholesale electricity markets are typically organized around the supply and demand for electrical energy and these markets yield time-varying spot prices for electrical energy. In the U.S., most of the organized spot wholesale markets for electrical energy place a cap on the price at which electrical energy can be sold. To the extent that these prices are binding, they are reflected as well in forward prices for energy in over-the-counter markets and bilateral transactions. These price caps have been put in place to mitigate market power in spot wholesale markets.

The organized wholesale markets in the Northeastern U.S. have also introduced "capacity

[21] Generating capacity may not always be available to supply energy due to planned maintenance and unplanned breakdowns or outages. Accordingly, the expected generating capacity available to supply electrical energy at a point in time is less then the nominal capacity. We ignore availability considerations in this section. Generating plants may also produce other products such as reactive power and frequency regulation which we also ignore here.

obligations" that are imposed on all LSEs. In these markets, LSEs not only have a financial obligation to purchase (or supply internally) all of the electrical energy consumed by their retail customers in real time, but also have an obligation to contract for sufficient generating capacity ex ante to meet their peak demand plus an administratively determined reserve margin reflecting uncertainties about future demand and generating capacity availability. Capacity obligations may take at least two forms. One requires LSEs to forward contract with generators to make their capacity available to the spot wholesale market operated by the system operator during peak demand periods, leaving the price for any energy supplied by this capacity to be determined ex post in the spot market. Alternatively, the capacity obligations could require forward contracting for both the price of capacity and the price of energy (or operating reserves) that this capacity supplies to the spot market during peak hours.[22]

For example, an LSE in the wholesale markets operated by the PJM Regional Transmission Organization covering states in the Mid-Atlantic and Mid-Western regions of the U.S. has an obligation to contract forward for generating capacity (but not the actual energy it may produce in real time) equal to its annual peak demand plus an administratively determined reserve margin. If it fails to have such contracts in place it is assessed a deficiency penalty. There is a market for capacity to meet LSE capacity obligations which yields a capacity price which we denote below as p_K and which we assume is less than the deficiency charge so the deficiency charge does not serve as a binding price cap on capacity prices.[23] We assume as well that the level of the capacity obligation selected by the regulator is optimal, reflecting consumer valuations for energy and the cost of blackouts. In what follows, we refer interchangeably to capacity obligations and the associated capacity prices p_K determined in the market for capacity. When the capacity obligations are placed on all LSEs (and thus on the peak demand placed on the system by all retail customers) it has the effect of requiring all demand to pay the capacity price and allows all qualifying generating capacity to receive the capacity price.

Proposition 1 shows that price-insensitive customers alone do not create a rationale for capacity obligations. Rather, there must be some reason why the spot price for energy does not

[22] 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.

[23] The capacity markets in the Northeast are being reformed in ways that yield an administratively determined price for capacity that varies with the system's actual reserve margin as measured by the ratio of generating capacity to peak demand. The formulas used to determine these prices are similar to those used between 1990 and 2001 in the old UK electricity pool to determine capacity prices. These formulas effectively serve as price caps on capacity prices as well. See Cramton and Stoft (2005).

fully adjust to reflect supply and demand conditions and differs from the correct economic signal. This section argues that capacity obligations and associated capacity prices have the potential to restore (partially or fully) investment incentives in a market in which generator market power or sheer regulatory opportunism induce regulators to keep spot prices low. It also briefly discusses other reasons why spot prices for energy may not provide the proper investment signals. Like in section 2.4, we specialize the model in most of this section to *two states of nature*.

The regulator may impose a price cap ($p_2 \leq p^{max}$) on wholesale energy prices, which in turn are reflected directly in retail prices given perfect competition among retailers, in order to prevent generators from exercising market power in the wholesale market during peak demand periods. Suppose that:

- baseload investment in generating capacity and production of energy is competitive (as earlier),
- peakload investment in generating capacity and production of energy 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. In a first step, we leave aside rationing; we later look at the two-way interaction between interruptability and the exercise of market power.

3.2 No rationing

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:

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

where p is, as earlier, the state-independent retail price. Letting $\widehat{\eta}_2 \equiv -\frac{\partial \widehat{D}_2}{\partial p_2}/\frac{\widehat{D}_2}{p_2}$ denote the elasticity of demand of the price-sensitive customers, one obtains the following Lerner formula:[24]

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

or,

$$p_2 = p_2^C, \text{ where } p_2^C \text{ is the Cournot price.}$$

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.*[25]

Condition (9), together with the market clearing equations ($K_1 = D_1(p) + \widehat{D}_1(p_1)$ and $K_1 + K_2 = D_2(p) + \widehat{D}_2(p_2)$) as well as the competitive retail price condition given an expected price p_2 at peak (as studied in section 2.4; for example, with linear demands, $p = f_1 p_1 + f_2 p_2 = p^*$), determines the equilibrium capacities, K_1 and K_2, and prices, p_1, p_2, p.

A price cap on energy production $p^{\max} = p_2^* = c_2 + (I_2/f_2)$ restores the Ramsey optimum. By contrast, a price cap creates a shortage of peakers whenever $p^{\max} < p_2^*$.[26]

Let us now show that (i) with two states of nature, the Ramsey optimum can nevertheless be attained through capacity obligations and the associated capacity prices even if the price cap on spot energy production is set too low, and (ii) with three states of nature, the combination of a price cap on energy production and capacity obligations and associated capacity prices 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

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

[24] 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

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

[25] 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.

[26] 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.

[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

$$p_K + f_2 p^{\max} = f_2 p_2^* \Longleftrightarrow p_K = f_2 \left(p_2^* - p^{\max}\right)$$

and so

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

so incentives for baseload production are unchanged, *provided that baseload plants are made eligible for capacity payments.*[27]

Another important point is that price-sensitive consumers consume too much (they consume $\widehat{D}_2 \left(p^{\max}\right)$ at peak) unless capacity obligations are imposed on LSEs in a way that reflects the peak demand of all of the retail consumers that they serve. 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. We thus conclude that price-sensitive consumers should be subject to capacity obligations.[28]

Proposition 4 Capacity obligations and the associated capacity prices 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 and that there are 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 in the ex post energy market through a combination of price cap on energy production and regulated capacity obligations and associated capacity prices, 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.

Remark: Regulators sometimes introduce price caps for reasons other than a desire to mitigate market power in the energy market. The following two examples discuss two of the primary

[27] 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.

[28] Similarly, the signal for penalizing a failure to deliver is lost: $p^{\max}$ is an underestimate of the social cost associated with a supplier's failure to deliver. The proper measure of the cost of default is therefore $p^{\max} + p_K$. The capacity obligation price p_K can also be used as the basis for computing the penalty for those LSEs that underpredict their peak demand and are short of capacity obligations.

additional rationales for imposing administrative price caps. Suppose that there is perfect competition and two states of nature but 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 peaking unit with the highest marginal operating 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:

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

The second best is then restored (subject to the caveats enunciated in the next 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).[29] 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.

Another rationale for a capacity obligation arises in the presence of a choke price: $\widehat{D}_2 (p_2^*) = 0$ (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 $D_2 (p^*) > K_2$ 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 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.[30]

[29] 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.

[30] Electricité de France (1994, 1995).

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.

3.3 Some limitations of capacity obligations

There are at least two *potential* problems that may result from a policy of applying binding price caps to the price of energy sold in the wholesale spot market that cannot be remedied with capacity obligations and associated capacity prices.

• *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, 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.[31]

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 solving

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

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

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

Proposition 5 With price-sensitive consumers, the combination of a price cap and capacity obligations has no impact on market power and the final allocation when generators rather than the regulator select the number of capacity obligations they supply.

[31] 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.

Note that we have not considered ex ante "bypass investments" by the consumers; anticipating high peak prices in the absence of price cap, consumers would overinvest in inefficient self-provision of peakers, putting downward pressure on the peak price until the spot price and the bypass cost are equalized. The study of bypass with or without price caps and with or without capacity obligations lies outside the scope of this paper.

Proposition 5 implies that the price or the quantity of capacity obligations must be regulated in order for the ex post price cap on energy production to alleviate market power.

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 (Allaz-Vila 1993).

• *A capacity payment is an insufficient instrument with more than two states of nature with market power.* 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 for any 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;[32] 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 $i \geq i_0$. 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:

$$p_K = E[(p^*_i - p^{\max})\, 1\!\!1_{i \geq i_0}]$$

[32] 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.

(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:

$$p_K > E[(p_i^* - p^{\max})\, \mathbb{1}_{i \geq k}].$$

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 $f_3\,(p_3^* - p^{\max}) = p_K$ implies that $f_2\,(p_2^* - p^{\max}) + f_3\,(p_3^* - p^{\max}) = p_K$.

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. Optimality can be achieved through an array of (state-contingent) capacity obligation markets, although one may worry about transaction costs and market power when markets multiply.

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, when unable to trade capacity obligations ex post, the consumer will overconsume for low idiosyncratic demand and underconsume 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).[33]

Proposition 6 With more than two states of nature with market power, a combination of a price cap and capacity obligations and associated capacity prices 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.

[33] 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.

3.4 Rationing

As section 2.4 showed, price-insensitive retail consumers are not rationed as long as the peak price is sufficiently low, i.e., if the cost of peakers is not too large and if there is enough competition among peakers to keep p_2 close to the long-term marginal cost of peakers.

It is worth considering the opposite case, in which the system operator or the LSEs would like the price-insensitive retail consumers to be interrupted during peak. Consider the case of a *monopoly* provider of peak generation (it is straightforward to extend the analysis to Cournot competition, as earlier). Fixing the retail price p and the off-peak capacity K_1 for the moment, consider the monopolist's choice of peak price p_2. It solves:

$$\max \left\{ \begin{array}{ll} \max & \left\{ (p_2 - p_2^*) \left[\widehat{D}_2 (p_2) + D_2(p) - K_1 \right] \right\}, \\ \{p_2 \leq \overline{p}_2(p)\} & \\ \max & \left\{ (p_2 - p_2^*) \left[\widehat{D}_2 (p_2) - K_1 \right] \right\}. \\ \{p_2 > \overline{p}_2(p)\} & \end{array} \right\}$$

That is, the monopolist can either charge a "low" price and keep the price-insensitive retail consumers on board, or charge a very high price and focus on the price-sensitive consumers.[34] The demand curve facing the monopolist is illustrated in figure 1.

INSERT FIGURE 1 ABOUT HERE

For expositional simplicity, let us assume that price-insensitive consumers' demand is linear ($D_i(p) = a_i - bp$) and so, in the absence of rationing,

$$p = p^* = E(p_i) = c_1 + I_1$$

(the analysis extends straightforwardly to non-linear demands). In all equilibria:

$$K_1 = \widehat{D}_1(p_1) + D_1(p)$$

and

$$f_1 p_1 + f_2 p_2 = c_1 + I_1.$$

There are three possible regimes (see figure 1):

• *Unconstrained, no-rationing equilibrium*: This is the regime studied in section 3.2 and depicted by point A in figure 1. The monopoly price, p_2^m, maximizes:

$$(p_2 - p_2^*) \left[\widehat{D}_2(p_2) + D_2(p^*) - K_1 \right]$$

[34] Note that we assume that the system operator has no commitment power and therefore optimizes over the interruptability decision ex post (and so rations retail consumers whenever $p_2 > \overline{p}_2(p)$). Otherwise, it could threaten to ration at a low peak price, so as to alleviate market power.

and satisfies:

$$p_2^m \leq \overline{p}_2(p^*).$$

Thus price-insensitive consumers are never rationed and are charged p^*.

• *Unconstrained, rationing equilibrium*: In this regime (depicted by point B in figure 1), the peaking monopolist charges $\widehat{p}_2^m$, which maximizes:

$$(p_2 - p_2^*)\left[\widehat{D}_2(p_2) - K_1\right].$$

Furthermore:

$$\widehat{p}_2^m > \overline{p}_2(p_1)$$

where p_1, the off-peak price (which is also the retail price), is given by the baseload units' free entry condition:

$$f_1 p_1 + f_2 \widehat{p}_2^m = c_1 + I_1 = p^*.$$

Because $\widehat{p}_2^m > \overline{p}_2(p_1) > p_1$, necessarily $p_1 < p^*$; hence $\overline{p}_2(p^*) < \overline{p}_2(p_1)$. As in the analysis of LSE competition and zonal rationing, the expectation that the consumer will be curtailed lowers the retail price, which makes the occurrence of rationing more likely.

• *Kinked, no-rationing equilibrium*: In this regime (depicted by point C in figure 1), the monopolist charges the highest price that is consistent with the price-insensitive consumers not being interrupted:

$$p_2 = \overline{p}_2\left(p^*\right).$$

This regime requires that $p_2^m > \overline{p}_2(p^*)$ and that the monopolist prefers to keep all consumers on board:

$$\max\left\{(p_2 - p_2^*)\left[\widehat{D}_2(p_2) - K_1\right]\right\} \leq [\overline{p}_2(p^*) - p_2^*]\left[\widehat{D}_2\left(\overline{p}_2(p^*)\right) + D_2(p^*) - K_1\right].$$

It can be shown that rationing and non-rationing equilibria may coexist. Intuitively, the absence of rationing leads to a high retail price, which, because $\overline{p}_2(\cdot)$ is increasing, allows the monopolist to demand a high peak price without triggering price-insensitive retail consumers'rationing; and conversely.

Proposition 7 When consumers are rationed optimally ex post,
(i) equilibrium behavior must belong to one of the three regimes described above;
(ii) there may be two equilibria under LSE competition, one with rationing and the other without.

4 Network collapses and operating reserves

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 is "orderly" in the sense that it 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 each for a relatively short time 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 total system 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). OpRes 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). OpRes 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 scheduled capacity K will break down. The distribution $H_i(\lambda)$ (with $H_i(0)=0$ and $H_i(1)=1$) can be state-contingent.[35] There may be an atom in the distribution at $\lambda = 1$ (full availability), but the distribution has otherwise a smooth density $h_i(\lambda)$.[36] 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.)

INSERT FIGURE 2 ABOUT HERE

[35] 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.

[36] 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.

The timing goes as in Figure 2: given nominal capacity K and demand D_i, the ISO 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.

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 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:

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

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

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

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

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:

$$\frac{h_i}{(1 + r_i)^2} v - s = \nu_i, \tag{10}$$

$$[1 - H_i]\, v - s\,(1 + r_i) = \mu_i + (1 + r_i)\, \nu_i, \quad \text{with } \mu_i = 0 \text{ unless } d_i = D_i \tag{11}$$

and

$$E\,[\nu_i] = I. \tag{12}$$

Specializing the model to the case in which H_i *is state-independent,*[37] let us analyze the optimal dispatching, as described by (10) and (11). The Ramsey optimum is depicted in Figure 3.

INSERT FIGURE 3 ABOUT HERE

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

Off-peak (D_i small), there is excess capacity, all demand is served ($d_i = D_i$), and $\nu_i = 0$. Hence from (10) and (11)

$$r = r_H$$

where[38]

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

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

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

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 keeps 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:

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

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 from a network collapse νD_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

[38] We assume that this equation has a unique solution, the equation $\lambda^2 h\,(\lambda) = s/\nu$ has a unique solution in $(D_i\ /K, 1)$. Because $\lambda h/\,(1-H)$ is inceasing, so is $\lambda^2 h\,(\lambda)$. And so the solution indeed exists as long as $\lambda_i h\,(\lambda_i) < s/\nu < h(1)$ for $\lambda_i = D_i\ /K$.

a reduction in load shedding, which has value $(1-H_i)\,v/(1+r_L)$ minus the cost s of scheduling generation.

Remark (adding price-sensitive consumers): A similar, although more complex analysis can be performed when price-sensitive consumers are present. Let $\widehat{S}_i\left(\widehat{D}_i\right)$ denote their gross surplus function in state i. The demand function $\widehat{D}_i(p_i)$, given by $\widehat{S}_i\left(\widehat{D}_i(p_i)\right) \equiv p_i$, is assumed to be increasing with i. There are again three possible regions: a no-capacity-constraint region, a capacity-constraint, no-load-shedding region, and a load-shedding region. In the former two regions, the price p_i charged to price-sensitive consumers and the reserve ratio r_i in general react to the state of demand, but price-sensitive consumers are not rationed, as $v > p_i$. In the third region, $p_i = v$ and price-insensitive consumers are rationed.[39]

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 full availability $\lambda = 1$. Thus, the market solution leads to an insufficient level of reliability. In order to obtain a proper level of reliability, the system operator must purchase (or require LSEs to purchase based on their retail customers' demand) a fraction r_i of reserves for each unit of load.[40]

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 $v/(1+r_L)$ conditionally on being actually served (which has probability $1-H_i$). Thus, generators obtain, as they should,

[39] The first-order conditions are:

$$\frac{h_i}{(1+r_i)^2}\left[v d_i + \widehat{S}_i\left(\widehat{D}_i\right)\right] = (s+\nu_i)\left(d_i + \widehat{D}_i\right),$$

$$\begin{aligned}(1-H_i)\,p_i &= (s+\nu_i)\,(1+r_i)\,,\\ (1-H_i)\,v &= (s+\nu_i)\,(1+r_i) + \mu_i\,.\end{aligned}$$

For an example, in the no-capacity-constraint region ($\nu_i = \mu_i = 0$),

$$\frac{h_i}{(1+r_i)\,(1-H_i)} = \frac{p_i\left(D_i + \widehat{D}_i\right)}{vD_i + \widehat{S}_i\left(\widehat{D}_i\right)}.$$

That is, the hazard rate of the reliability distribution is equal to the ratio of the market value of demand to the corresponding social surplus.

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

quasi-rent:

$$(1 - H_i)\,\frac{\nu}{1 + r_L} - s$$

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 $\nu/(1 + r_i)$ 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, $(h_i \frac{D_i}{K^2}) \cdot (D_i \nu) - s$, 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:

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

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

Proposition 8 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 price-insensitive 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.

Another lesson of our analysis is the possibility of self-fulfilling rationing prophecies. Consumers are interdependent when either the system operator must ration at the zonal rather than individual level or when a generator with market power chooses its price. The system operator in the former case, and the generator in the latter, base their choice on the average retail contract. LSEs charge a low price to their customers if they expect these to be rationned, which generates very large peak demands and ex post socially optimal rationning of these customers.

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 (2005) 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. and Vila, J.L. "Cournot Competition, Forward Markets and Efficiency." *Journal of Economic Theory*, Vol. 59 (1993), pp. 1–16.

[2] Borenstein, S. and Holland, S. "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices." 2003. Forthcoming, *Rand Journal of Economics*.

[3] ——, Jaske, M., and Rosenfeld, A. "Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets." Center for the Study of Energy Markets. Paper CSEMWP-105, October 2002.
http://repositories.cdlib.org/ucei/csem/CSEMWP-105.

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

[5] —— and —— "Resource Adequacy and Market Power Mitigation Via Option Contracts." Mimeo, (December 2003). Available as Chapter 1 of "Implementation of Resource Adequacy Requirements via Option Contracts," Report 1010712, Electric Power Research Institute, October 2005.

[6] Cramton, P. and S. Stoft "A Capacity Market that Makes Sense," Working Paper University of Maryland, mimeo, March 2005. http://www.cramton.umd.edu/papers2005-2009.

[7] Electricité de France "The Explicit Cost of Failure." Mimeo, General Economic Studies Department, 1994.

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

[9] Joskow, P. and Tirole, J. "Retail Electricity Competition." Mimeo, MIT and IDEI, 2005. Forthcoming, *Rand Journal of Economics*.

[10] —— and —— "Procurement by the ISO: Supplementary Material for "Reliability and Competitive Electricity Markets"." Mimeo MIT and IDEI, 2005.

[11] Littlechild, S. "Why We Need Electricity Retailers: A Reply to Joskow on Wholesale Spot Price Pass-Through." Working Paper no. 0008, Department of Applied Economics, University of Cambridge, 2000.

[12] Oren, S. "Ensuring Generation Adequacy in Competitive Electricity Markets." Mimeo, UC Berkeley, 2003.

[13] Stoft, S. *Power System Economics*. New York: Wiley, 2002.

[14] —— "The Demand for Operating Reserves: Key to Price Spikes and Investment." *IEEE Transactions on Power Systems*, Vol. 18 (2003), pp. 470–478. .

[15] Turvey, R."Profiling: A New Suggestion." Mimeo, 2003.

[16] —— and Anderson, D. *Electricity Economics: Essays ands Case Studies*. Baltimore and London: Johns Hopkins University Press, 1977.

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

[18] U.S.- Canada Power System Outage Task Force. "Interim Report: Causes of the August 14 Blackout in the United States and Canada." November 19, 2003.

[19] Vasquez, C., Rivier, M., and Perez-Arriaga, I. "A Market Approach to Long-Term Security of Supply." *IEEE Transactions on Power Systems*. Vol. 17 (2002), pp. 349–357.

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

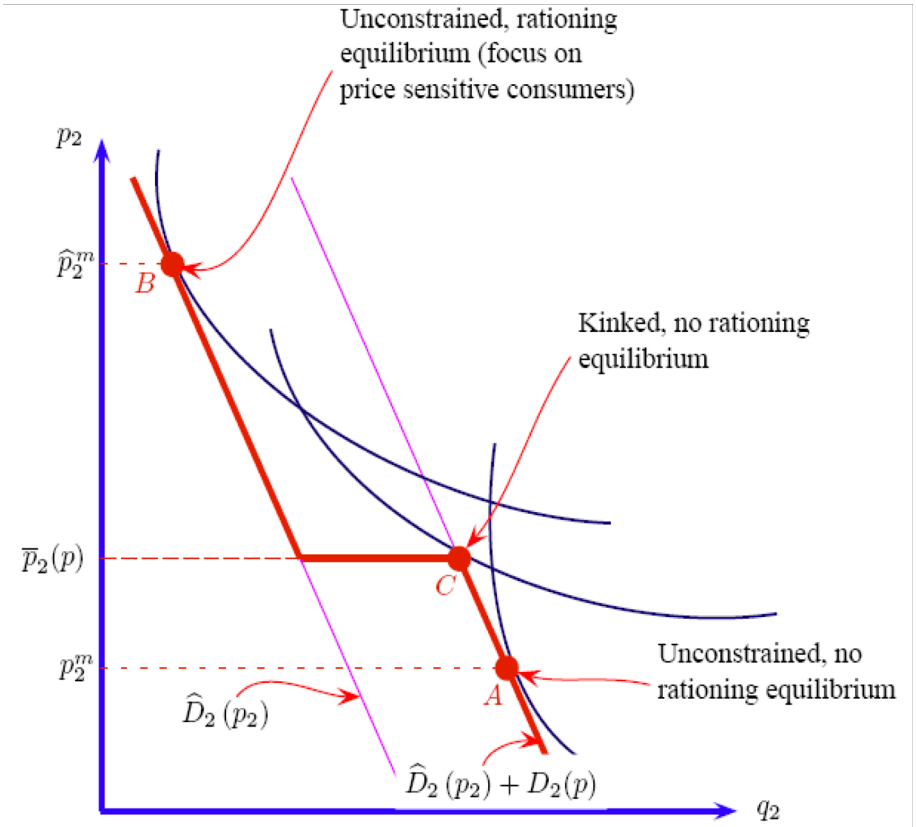

Figure 1: **Demand for price-sensitive and price-insensitive retail consumers**

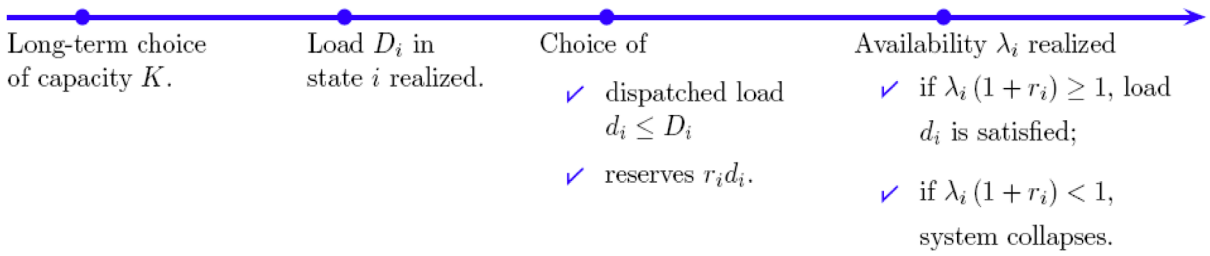

Figure 2: Timing of capacity commitment, dispatch and load shedding

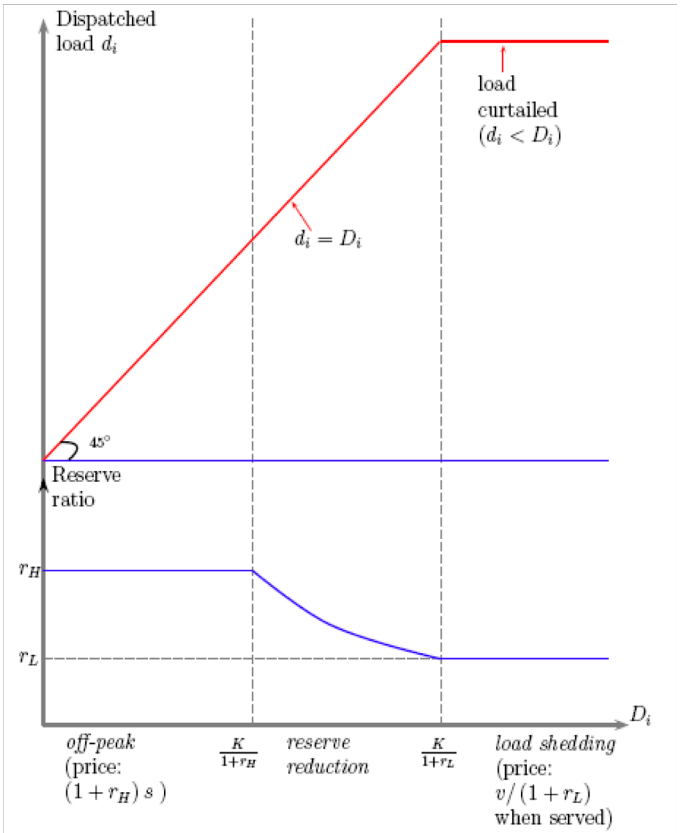

Figure 3: Ramsey optimum (prices indicated in parentheses are prices paid by consumers)