Reliability and Competitive Electricity Markets

Paul Joskow* and Jean Tirole†

January 12, 2004

Very preliminary. Please do not circulate. Comments are welcome. [These notes have been put together for the IDEI-CEPR conference on "Competition and Coordination in the Electricity Industry," January 16–17, 2004, Toulouse.]

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

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

Abstract

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. Much of the 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. The paper is an attempt at bridging the gap between the economists' and the engineers' approaches.

We build upon the relaxation of the four key assumptions underlying the following benchmark proposition: even in an environment with price-insensitive consumers and rationing, the second best optimum can be implemented by an equilibrium with retail and generation (wholesale) competition provided that: (a) Load Serving Entities (LSEs) face the real time price for the aggregate consumption of the retail customers for whom they are responsible. (b) The real time wholesale price accurately reflects the social opportunity cost of generation. (c) Rationing, if any, is orderly, and makes use of available generation. (d) Consumers who can react fully to the real time price are not rationed. Furthermore, the LSE 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.

The paper first derives the optimal prices and investment program when there is uncertain demand, consumers who do not react to real time prices, and price rationing of consumers to balance supply and demand in high demand states. This leads to the benchmark decentralization proposition summarized above. It then analyzes the implications of load profiling for retail competition. Third, in a situation in which either generator market power or regulatory opportunism distort wholesale prices, it studies whether capacity obligations and/or purchases of peaking capacity by the system operator can provide appropriate investment incentives as well as efficient spot markets. Fourth, it derives the implications of network collapses and the concomitant need of network support services. It argues that network collapses differ from other forms of energy shortages and rationing in a fundamental way, and discusses the implementation of the Ramsey allocation through a combination of regulation and market mechanisms. Finally, the paper analyzes the implications of limitations in the controllability of the distribution circuits; it discusses both market mechanisms that are needed to reach a "third best" and the difficulties that make the phasing out of non-market mechanisms unlikely.

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: default retail service obligations placed on incumbent distributors, wholesale market price caps, capacity obligations placed on LSEs, frequency regulation, operating reserve and other ancillary service requirements enforced by the system operator, procurement obligations placed on system operators, protocols for non-price rationing of demand to respond to "shortages", and administrative protocols for system operators' management of system emergencies. Many of these non-market mechanisms have been carried over from the old regulated regime without much consideration of whether and how they might be replaced with market mechanisms and of the effects they may have on market behavior and performance if they are not.

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

Other mechanisms and requirements have been justified by what are perceived to be special physical characteristics of electricity and electric power networks which in turn lead to market failures that are unique to electricity. These include the need to meet specific physical criteria governing network frequency, voltage and stability that are thought to have public good at-

tributes, the rapid speed with which responses to unanticipated failures of generating and transmission equipment must be accomplished to continue to meet these physical network attributes and the possibility that market mechanisms cannot respond fast enough to achieve the network's physical operating parameters under all states of nature.

Much of the 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 institutional environment in which our analysis proceeds has competing load serving entities (LSEs)[1] that market electricity to residential, commercial and industrial ("retail") consumers. LSEs may be independent entities that purchase delivery services from unaffiliated transmission and distribution utilities or they may be affiliates of these transmission and distribution utilities that compete with unaffiliated LSEs. Some retail consumers served by LSEs can respond to real time wholesale market prices while others cannot. Some retail consumers' consumption can be measured in real time, even though they may not have the capability to respond to real time prices, while others' cannot and their real time consumption must be estimated using load profiling procedures. Retail consumers may be subject to non-price rationing to balance supply and demand in real time. The wholesale market is composed of competing generators who compete to sell power to LSEs.

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

The wholesale market may be perfectly competitive or characterized by market power. Finally, there is an independent system operator (ISO) which is responsible for operating the transmission network in real time to support the wholesale and retail markets for power, including meeting certain network reliability and wholesale market power mitigation criteria.[2]

Section 2 derives the optimal prices and investment program when there is uncertain demand, consumers who do not react to the real-time price and non-price rationing of consumers to balance supply and demand during very high demand states. 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 some price-insensitive consumers) can be implemented by an equilibrium with retail and generation (wholesale) competition provided that:

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

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

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

(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.

[2]The latter may include enforcing operating reserve and other operating reliability requirements, enforcing longer term capacity obligations, procuring and dispatching resources to meet these requirements, and managing system emergencies that might lead to network to collapse.

The assumptions underlying this benchmark proposition are obviously very strong: (a) LSEs do not face the real time price for their customers if these customers are load profiled; (b) market power on the one hand, and price caps and other policy interventions on the other hand create differences between the real time wholesale market price and the social opportunity cost of generation; (c) network collapses, unlike say rolling blackouts, have systemic consequences, in that available generation cannot be used to satisfy load; (d) price sensitive consumers may be rationed along with everyone else that is physically connected to the same controllable distribution circuit; and, relatedly, LSEs generally cannot demand any level of rationing they desire. Accordingly, the paper then proceeds in the remaining four sections to examine the implications of relaxing these assumptions.

Section 3 analyzes retail competition between LSEs in a world in which consumers differ in the metering equipment that they have installed, the associated variations in the capability to measure their actual real time consumption, and their responsiveness to real time prices. We build on the analysis of Borenstein and Holland (2003a,b) and expand on it in a number of ways: (i) we specifically take load profiling into account and distinguish between price-insensitive consumers whose real time consumption can be measured and those whose real time consumption must be estimated through load profiles; (ii) we allow competing LSEs to offer non-linear prices to retail consumers; (iii) we consider regimes in which the incumbent distributor can also compete for retail consumers as an LSE and those in which only independent retailers are allowed to compete in the retail market. We show that under load profiling, retail competition (with or without the incumbent distributor) leads to a retail price equal to the average wholesale power cost and differing from the socially optimal retail price. We show on the other hand, that given

the price inefficiencies associated with retail competition, LSEs face the right incentives when offering their customers enhanced metering equipment.

Section 4 studies the implications of distorted wholesale prices. It first considers the case where there is a competitive supply of base load generation, market power in the supply of peak load investment and production, and a price cap is applied that constrains the wholesale market price to be lower than the competitive price during peak periods (section 4.1). This creates a shortage of peaking capacity in the long run when there is market power in the supply of peaking capacity. We show that capacity obligations and associated capacity prices have the potential to restore investment incentives by compensating generators ex ante for the shortfall in earnings that that they will incur due to the price cap. Indeed, with up to three states of nature, the Ramsey optimum can be achieved despite the presence of market power through a combination of a price cap and capacity obligations provided that : (i) both peak and base load generating capacity are eligible to meet LSE capacity obligations and receive the associated capacity price, (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, and (iii) the market for peaking capacity is contestable (which is likely to require capacity obligations a few years ahead). With more than three states of nature, a combination of spot wholesale market price caps and capacity obligations will not achieve the Ramsey optimum unless market power is only a problem during peak demand periods. Thus, the regulator faces a tradeoff between alleviating market power off-peak, if it is a problem, through a strict price cap, and providing the proper peak investment incentives, and is further unable to provide price-sensitive consumers with the appropriate economic signals.

Section 4.2 then examines the effects of two types of behavior by an ISO that empirical analysis has suggested may distort prices and investment (Patton 2002). The first involves inefficient dispatch of resources procured by the ISO for use during operating reserve deficiency conditions. Such dispatch in the short run depresses off-peak prices and in the long term leads to an inefficient substitution of base load units by peakers. The second involves the recovery of the costs of resources acquired by the ISO through an uplift charge spread over prices in all demand states or else in only peak demand states. Whether the uplift is socialized (spread over demand states) or not, large ISO purchases discourage the build up of baseload capacity and depresses the peak price. For small purchases, off-peak capacity decreases under a socialized uplift, and peak capacity decreases under an uplift that applies solely to peak energy consumption.

Section 5 derives the implications of network collapses and the concomitant need for network support services. As suggested above, network collapses differ from other forms of energy shortages and rationing in a fundamental way. While scarcity makes available generation (extremely) valuable under orderly rationing, it makes it valueless when the network collapses. 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 system services, and discuss the implementation of the Ramsey allocation through a combination of regulation and market mechanisms.

Last, section 6 analyzes the implications of limitations in the controllability of the distribution circuits. These limitations imply that price sensitive consumers may be rationed along with everyone else, and that LSEs cannot generally demand any level of rationing that they desire. At best one can

then elicit only the aggregate willingness to pay for reliability in any given joint interruptibility zone. The section discusses both market mechanisms that are needed to reach a "third best" and the difficulties that make the phasing out of non-market mechanisms unlikely.

2 A benchmark decentralization result

In order to later analyze competition among LSEs for the final (retail) consumers, it is convenient to group the latter into three categories:[3]

- *Price-sensitive consumers* are endowed with real-time (RT) meters and either autonomously or through communication with the LSE, adjust their demand to the evolution of the wholesale spot market price.

- *Price-insensitive consumers with recording demand meters* are endowed with RT meters, but are only partially aware or unaware of RT prices and therefore do not adjust their consumption perfectly as real time prices vary from minute to minute and hour to hour. At the extreme, they are fully (RT) price-insensitive. While consumers do not react to real time prices their actual real time consumption can be measured and assigned to their LSE for settlement purposes.

- *Load-profiled consumers* have traditional meters. They are metered only once a month or every few months (e.g. in some countries meters are read even less frequently), and pay a per-kWh electricity charge that is independent of the actual timing of their overall consumption.

[3]The grouping in three categories is an oversimplification. There are a number of partially price sensitive categories, such as those subject to time-of-use pricing (retail prices are preset for certain blocks of time) or critical peak pricing (that combines time-of-use pricing with high retail prices for a number of critical hours per year to be declared by the utility). See Borenstein et al (2002) for a review of recent innovations.

Unlike in the previous case, in which the LSE serving the consumer faces the RT wholesale price, an LSE other than the local distribution grid owner and serving such a consumer pays a unit electricity charge based on the "load profile" of the consumer. That is, it pays *the average wholesale price for the load profile* that is representative of the consumer's class regardless of the actual time patterns of the individual customer's consumption and the relationship between this actual physical consumption and the contemporaneous RT wholesale prices.

Meter Type	**Consumption Measurement**	**Customer RTP Sensitivity**
Traditional	Load profile	No
Recording Demand	Actual Real Time	No/partial
Real Time Pricing	Actual Real Time	Yes

TYPES OF RETAIL CONSUMERS

Table 1

We leave the analysis of load profiling to section 3, and assume that LSEs face the real time wholesale price for the aggregate consumption of the retail customers for whom they are responsible.

2.1 Model[4]

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

Price-insensitive consumers do not react to the RTP. They pay a constant

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

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

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 (say, determined by the system operator); alternatively, one could envision situations in which the LSEs would affect the alphas either by demanding that their consumers not be served as the wholesale price reaches a certain level, or conversely by bidding for priority in situations of rationing.[6] We let $\mathcal{D}_i(p, \alpha_i)$ denote their expected consumption in that state, and $\mathcal{S}_i(p, \alpha_i)$ their realized gross surplus, with

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

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

In the *separable case*, the demand $\mathcal{D}_i$ takes the multiplicative form $\alpha_i D_i(p)$ and the surplus takes the separable form $\mathcal{S}_i(D_i(p), \alpha_i)$. More generally however, the consumer may adjust her demand to the prospect of being potentially rationed.[7]

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 $S_i(D_i(p)) - pD_i(p)$.

[6] 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.

[7] 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.

Let us now discuss specific cases to make this general 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.[8]

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 $S_i\left(D_i\left(p\right)\right)$), 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

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

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 and gets zero surplus from the elevator. By contrast, the consumer obtains a negative surplus from the elevator if the outage is unforeseen. Similarly, consumers would have planned an activity requiring no use of electricity (going to the beach rather than using the washing machine, drive their car or ride their bicycle rather than use the subway) if they had anticipated the blackout; workers could have planned time off, etc. More generally, with adequate warning consumers can take advance actions to adapt to the consequences of an interruption in electricity supplies. This is one reason why distribution companies notify consumers about planned outages required for maintenance of distribution equipment.

Opportunity cost example: Suppose that the consumer chooses between an electricity-consuming activity (taking the elevator, using electricity to run

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

an equipment) and an electricity-free approach (taking the stairs, using gas to run the equipment). The latter yields known surplus $\overline{S}$. The surplus associated with the former depends not only on the marginal price p he faces for electricity, but also on the probability $1-\alpha$ 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 $S_i^n(p) \equiv \max\{S_i(D) - pD\}$ 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, in case (c), that $E[\alpha_i]$ is high enough so that the consumer chooses the electricity-intensive approach; and that $S_i^n(p) \geq \overline{S}$).

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

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

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

When $\mathcal{D}_i = \alpha_i D_i$, then

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

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

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

The supply side is described as a continuum of investment opportunities indexed by the marginal cost of production c. Let $I(c)$ denote the investment cost of a plant producing one unit of electricity at marginal cost c. There are constant returns to scale for each technology. We denote by $G(c) \geq 0$ the cumulative distribution function of plants.[9] So, the total investment cost is

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

The ex post production cost is

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

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

where the utilisation rate $u_i(c)$ is 1 for $c < c_i$, and 0 for $c > c_i$ for some c_i.

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 $H_i(\lambda)$). This would not alter the conclusions.

2.2 Optimum and competitive equilibrium

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

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

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

a) *Efficient dispatching*:

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

b) *Price-sensitive consumers*:

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

c) *Price-insensitive consumers*:

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

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

d) *Investment*:

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

These first-order conditions can be interpreted in the following way: condition (1) says that only plants whose marginal cost is smaller than the dual price p_i are dispatched in state i. Condition (2) implies that price-sensitive consumers are never rationed and that their consumption decisions are guided by the state-contingent dual price.[10] Condition (3) yields a covariance formula for the price $p = p^*$ provided that price-insensitive consumers are never rationed ($\alpha_i \equiv 1$):

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

We will return to this formula in section 3.

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

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

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

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

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

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

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

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

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

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

By contrast, with imperfectly foreseen outages,

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

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

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

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

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

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

The following proposition shows that rationing does not alter the conclusions that retail competition is consistent with Ramsey optimality in the absence of load profiling, provided that four assumptions are satisfied:

Proposition 1 *The second-best optimum (that is, the socially optimal allocation given the constraint that retail consumers do not react to the RTP, but their real time consumption can be measured) can be implemented by an equilibrium with retail and generation competition provided that:*

- *load-serving entities face the RTP,*
- *price-sensitive consumers are not rationed; furthermore, while price-insensitive consumers may be rationed, their load-serving entity can demand any level of state-contingent rationing* $\alpha_i(p_i)$,[11]
- *the RTP reflects the social opportunity cost of generation,*
- *available generation is made use of during rationing periods.*

Proof: Suppose that retailers can offer 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:

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

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

[11] 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 has long as LSEs can select a state-contingent α_i.

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

2.3 Two-state example

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

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

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

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

where

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

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

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

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

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

and

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

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

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

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

Intuitively, for infrequent peaks, the peak price goes to infinity and so the discrepancy between the true price and the price paid by retail consumers is too large to make it socially optimal to serve the consumers.

3 Retail competition and load profiling

As retail competition for price-insensitive consumers with recording demand meters and price-sensitive ones is studied in Borenstein-Holland (2003a,b), who ignore load profiling, this section's primary goal is to analyze retail competition among LSEs for consumers subject to load profiling. We derive two benchmark allocations for cases where consumers are not sensitive to the real time price. The first characterizes the second-best (Ramsey) two-part tariff when retail consumers do not see or react fully to real time prices. The second characterizes the competitive equilibrium under retail competition when all consumers' real time demand can be measured but they cannot see or react to real time prices. Next we proceed to analyze the allocations that result when there is retail competition for load profiled consumers only and compare them to these benchmarks. Finally, we examine consumer/LSE incentives for load profiled customers to install two different types of more advanced metering equipment that either allow real time consumption to be measured or allow both measurement of real time consumption and reactivity to real time prices.

States of nature (or, equivalently, periods) are again indexed by $i \in [0, 1]$. f_i denotes the frequency of state i. Because this section focuses on competition on the demand side, we take the wholesale prices as exogenous, and we identify states of nature by the wholesale price p_i, with p_i increasing in i.

For the sake of simplicity, let us ignore rationing in this section ($\alpha_i \equiv 1$). We consider a representative retail consumer with demand $D_i(p)$ when facing

price p in state i, with $D_i' < 0$. Let $S_i\left(D_i(p)\right)$ denote the associated gross surplus, with $S_i' = p$. Note that consumers are assumed to be homogeneous (they may differ in the size of their demand, though: That is, they can be indexed by $\sigma > 0$, such that a consumer of type σ has demand $\sigma D_i(p)$ and surplus $\sigma S_i\left(D_i(p)/\sigma\right)$. More general forms of heterogeneity are briefly analyzed in Appendix 1).

Assumption 1. The function $E\left[\left(p - p_i\right) D_i'(p)\right]$ is decreasing in p.

The retail consumers are physically served by a local grid owner (usually also called the incumbent distributor, or transmission and distribution service provider). Because we are not interested here in the price of access to the grid, we normalize to zero any delivery, metering and customer service costs that continue to reflect responsibilities of the distribution grid owner. Thus, the LSEs' only cost is either the RT purchase of energy from the pool or, in the case of load-profiled consumers, the load profiled variable charge to be paid for power delivered by the local grid owner.[12]

[12]In this formulation, the aggregate demand of all consumers served through a particular distribution network is measured on a real time basis. However, individual customers may or may not have their consumption measured on a real time basis. Individual retailers may have a mix of customers with and without real time meters. Since the aggregate real time consumption obligations must add up to the aggregate real time supplies of power delivered over the distribution network, a set of "load profiles" must be applied to the monthly, bi-monthly or quarterly consumption measured for customers without real time meters. For example, consider a customer with a standard meter read on a monthly basis with 1000 kWh of consumption recorder for the previous month. The 1000 kWh of monthly consumption then must be allocated to the 720 hours of the previous month for settlement purposes. This is accomplished by assigning the customer to a group or class of customers thought to have similar consumption. A consumption or load profile is developed for each group based on real time metered consumption patterns of a sample of customers in each class. An individual customer who consumed no electricity during very hot summer days (because she was on vacation for half the month) would still have her measured monthly consumption allocated to some hot summer day hours based on her group's load profile. The load profile-based allocations must also satisfy an adding up property so that all power measured to have flowed through the distribution network is fully allocated to retail consumers. There are at least two ways to do this. One way is to load profile all customers without real time meters whether they are served by competitive retailers or the distribution company providing default retail service. Another

We allow LSEs (retailers) to offer to consumers two-part tariffs, consisting of a monthly subscriber charge and a per-kWh variable charge.[13] We analyze the competitive outcome in two environments. In the first, the local grid owner is subject to a line-of-business restriction. He provides access or delivery service to retailers, but is not allowed to compete for the final consumer. In the second, this line-of-business restriction is lifted and so the incumbent distributor is permitted to compete with independent retailers. We assume either that the distributor separates its retail "supply" business into a ring-fenced affiliate that is treated like any other retailer (as in the UK and in Texas), or that the retail arm maximizes the profit of the vertically integrated firm.[14] Before we study these two environments, it is useful to define two benchmark cases that are directly relevant to the cases studied carefully in Borenstein and Holland (2003a,b).

3.1 Two benchmarks

a) *Ramsey optimum for price insensitive consumers*

A non-responsive consumer cannot obtain the first-best utility, U^{FB}, that she would obtain if her demand were controlled to perfectly adjust to the RTP:

$$U^{FB} \equiv E\left[S_i\left(D_i\left(p_i\right)\right) - p_i D_i\left(p_i\right)\right]. \tag{9}$$

way is to load profile only the customers with traditional meters of competitive retailers and subtract the resulting hourly aggregates from the real time metered consumption for the entire distribution system, leaving the distribution company/retailer with settlement obligations for the residual.

[13] Offers by retailers to residential customers in England and Texas that we have reviewed have a fixed monthly charge plus one or more tiers of kWh charges.

[14] A further complication is that when retail competition is first introduced the distributor as retailer initially cannot "compete" in the normal sense, but rather is required to offer default service at a regulated price. These default service prices have been set in many different ways. We view these regulated default service obligations as transition arrangements and focus our analysis on a post transition retail competition regime where there is no regulated default service requirement.

A Ramsey social planner for consumers with traditional or recording demand meters who cannot respond to real time prices would choose prices, namely single per unit retail price p^* and fixed fee A^*, so as to maximize the consumer's expected net surplus subject to the budget balance constraint:

$$\begin{aligned} U^* \quad &\equiv \quad \max_{\{p^*, A^*\}} E\left[S_i\left(D_i\left(p^*\right)\right) - p^* D_i\left(p^*\right)\right] - A^* \\ &\text{s.t.} \\ &E\left[\left(p^* - p_i\right) D_i\left(p^*\right)\right] + A^* \geq 0. \end{aligned} \tag{10}$$

At the optimum, the budget constraint is binding, and the Ramsey planner maximizes the joint surplus:

$$U^* = \max_{p^*} E\left[S_i\left(D_i\left(p^*\right)\right) - p_i D_i\left(p^*\right)\right], \tag{11}$$

yielding the following covariance formula:

$$E\left[\left(p^* - p_i\right) D_i'\left(p^*\right)\right] = 0. \tag{12}$$

Assumption A1 implies that (12) has a unique solution.

To get some feel for what the Ramsey price entails, suppose for example that the elasticity of demand comes from the installation of air conditioning units. Suppose further that there are only two periods: off-peak (1) and peak (2), with respective wholesale prices p_1 and p_2. Then, the Ramsey price is $p^* = p_2$ in the US and $p^* = p_1$ in France, since summer is part of the peak in the US and is off peak in France.[15] Thus the Ramsey price would be greater than the average annual wholesale price of electricity in the US and below the average annual wholesale price in France.

Remark (optimality of two-part tariffs): We have assumed that the Ramsey

[15] If "installations" referred to electric heating, then $p^* = p_2$ in France since winter is the peak period.

planner offers two-part tariffs. Could a better allocation be obtained through more complex pricing structures?

With *traditional meters*, the social planner cannot do better than with a two-part tariff. At best he can hope to control total consumption through the marginal charge, while the load curve is chosen by the consumer without any concern for the actual cost of purchasing energy. More formally, the social planner is limited to total-consumption based tariffs $T(Q)$. Suppose that the planner selects the consumer's total consumption Q, and charges an amount T for this. The consumer then chooses her load curve so as to solve:

$$\max E\left[S_i\left(D_i\right)\right] \quad \text{subject to} \quad E\left[D_i\right] = Q.$$

Letting p denote the shadow price of the constraint, $S_i'\left(D_i\right) = p$, and so the allocation is the same as under a two-part tariff. By contrast, with *recording demand meters*, the social planner may or may not be able to do better than

with a two-part tariff.[16]

b) *Competition between retailers for price-insensitive consumers equipped with recording demand meters.*

Let us next assume that perfectly competitive (Bertrand) retailers compete for non-load-profiled price-insensitive consumers. The key point here is that the LSE's customers' actual consumption in each interval can be measured for billing and settlement purposes (that is, there is no need to load profile) but they don't see the prices.

With *two-part tariffs*, LSEs' attempt to woe the consumer leads them to solve program (10) and thus to offer the Ramsey optimum.

When retailers are constrained to offer *linear* prices, though, Bertrand

[16] An unrealistic but illuminating case is the following: Suppose that all price fluctuations are due to demand fluctuations, and that the consumer perfectly controls / is aware of demand charges (more on this shortly). Then, the social planner can punish the customer for consumptions above or under $D_i(p_i)$ in state i, and thereby obtain U^{FB} instead of the second-best level U^*.
Two equally unrealistic, but polar examples in which the consumer does not monitor the RTP and the social planner cannot obtain more than U^* despite the presence of recording demand meters go as follows. Suppose, first, that all price fluctuations are due to unavailability of plants or breakdowns of transmission lines. The consumer's demand is $D_i(p) = D(p)$. The optimal Ramsey price (given by (12)) is then constant: $p^* = E[p_i]$. Another, perhaps more interesting, example is supplied by the choice of an equipment (heater, air conditioning, pool) that, for a given quality of service s (set once and for all by the consumer) consumes a state-contingent amount of electricity $D_i(s)$. The marginal price of electricity affects the quality $s(p^*)$ (for example, an increase in p^* raises the temperature chosen by the consumer or induces the consumer to switch to oil heat), but to the extent that the consumer does not adjust the settings in a state-contingent way, recording demand meters do not improve on traditional ones.
In general, though, one could expect *some* reactivity of consumers to the RTP under recording demand metering. To be certain, this reactivity to the RTP would come at a price: Consumers would incur transaction costs in monitoring (at least from time to time or unconsciously) the state of nature and/or the RTP in order to adjust their consumption. Those transaction costs per se are no argument against using retail tariffs based on the RTP: The "true prices" are offered to the consumer, who then decides how much attention to pay to them. A potential argument against the use of RTP for consumers with recording demand meters is that it would obfuscate price comparisons with existing tariffs; however, websites already facilitate such price comparisons in the case of consumers with traditional meters. Another potential argument against RTP relates to the consumers' solvency or risk aversion; LSEs however could bundle small-scale "contracts for differences" with their supply contracts for consumers with recording demand meters.

competition drives the per-kWh price down to the "average wholesale cost price" $\hat{p}$, given by

$$E\left[\left(\hat{p}-p_i\right)D_i\left(\hat{p}\right)\right]=0. \tag{13}$$

That is, $\hat{p}$ is equal to the average wholesale price of the electricity purchased by the LSE to serve its customers. Borenstein and Holland (2003a,b) study the relationship between the Ramsey price per unit p^* and $\hat{p}$. Intuitively, $\hat{p}$ exceeds p^* if the state of nature impacts demand more than marginal demand.

Comparing (12) and (13), we are thus led to consider three cases:

Case 1: $\dfrac{E\left[p_iD_i(p)\right]}{E\left[D_i(p)\right]} > \dfrac{E\left[p_iD_i'(p)\right]}{E\left[D_i'(p)\right]}.$ for all p.

In this case, $A^*>0$ and $p^*<\hat{p}$.

Case 2: $\dfrac{E\left[p_iD_i(p)\right]}{E\left[D_i(p)\right]} < \dfrac{E\left[p_iD_i'(p)\right]}{E\left[D_i'(p)\right]}.$ for all p.

In this case, $A^*<0$ and $p^*>\hat{p}$.

Case 3: $\dfrac{E\left[p_iD_i(p)\right]}{E\left[D_i(p)\right]} = \dfrac{E\left[p_iD_i'(p)\right]}{E\left[D_i'(p)\right]}.$ for all p.

In this case, $A^*=0$ and $p^*=\hat{p}$.

Examples: For the additive linear with state-contingent intercept case $D_i(p)=d_i-h(p)$, we are in case 1. For the multiplicative case, $D_i(p)=d_ih(p)$, then $A^*=0$ and $p^*=\hat{p}$ (case 3).

3.2 Retail competition for load-profiled consumers: independent retailers

Next suppose that (pure) retailers, but not the local grid owner, compete for load-profiled consumers.[17] Consumers' real time demand is recorded by traditional meters and they do not see variations in real time prices so they cannot respond to them. Retailers' settlement obligations for wholesale power costs are then based on their customers' load-profiled consumption. To compute the price per kWh paid for wholesale energy by retailers for each customer they have signed up, a, suppose that, in equilibrium, retailers' per-kWh charge to consumers is p. Average consumption per consumer is $E[D_i(p)]$ and the wholesale price paid by the retailers for energy is

$$a(p) = \frac{E[p_i D_i(p)]}{E[D_i(p)]}. \tag{14}$$

We use the notation a for "access charge" by analogy with the economics literature on variable charges paid by entrants for access to regulated bottlenecks (local loop, etc.).[18] This access charge must be understood as the average wholesale power cost paid by retailers.

The following characterization will prove useful later on:

Lemma 1. (i) *Cases 1 through 3 can be characterized by how the average wholesale cost price varies with the marginal retail prices:*

[17] We are interested solely in the price effects of retail competition. We thereby ignore some benefits of competition(such as improved incentives to offer better metering, tariffs, total energy management services or hedging packages) as well as some potential costs of retail competition (such as consumer churn and poaching, duplicative or misleading advertising expenditures, and competitive screening for credit quality and high volume consumers).

[18] Note that our setup is equivalent to assuming that the distribution grid owner purchases the power in the wholesale market and then resells it to each LSE based on the real time metered or load profiled consumption of the customers they have signed up. The access charge a is then the price LSEs pay to compensate the distribution grid for the costs of the wholesale power it has purchased on their behalf.

$a' > 0$ *in case 1*

$a' < 0$ *in case 2*

$a' = 0$ *in case 3.*

(ii) *In all cases,* $a(p) > p$ *for* $p < \widehat{p}$

$a(p) < p$ *for* $p > \widehat{p}$.

Proof: Part (i) is obtained by deriving (14). To demonstrate part (ii), it suffices to show that $a'(p) < 1$ whenever $a(p) = p$, or after a few computations:

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

We know that $a(p) > p$ for p small (since $a(p) \geq E[p_i]$) and $a(p) \leq p_1 < p$ for p going to infinity. Hence, if the equation $a(p) = p$ has multiple solutions (an odd number greater than one) the function $H(p)$ must be increasing over at least some range. But $H'(p) = E\left[2D_i' + (p - p_i) D_i''\right] < E\left[D_i' + (p - p_i) D_i''\right] < 0$, a contradiction. ■

A retailer designs his offers so as to solve:

$$\max_{\{p,A\}} E\left[(p - a) D_i(p)\right] + A$$

s.t.

$$E\left[S_i\left(D_i(p)\right) - pD_i(p)\right] - A \geq \bar{U},$$

where $\overline{U}$ is the net surplus obtained by the consumer from subscribing with a rival retailer.

The retailer therefore selects p so as to maximize the joint surplus:

$$\max_p E\left[S_i\left(D_i(p)\right) - aD_i(p)\right],$$

or

$$(p - a) E\left[D_i'(p)\right] = 0,$$

yielding

$$p = a.$$

In equilibrium, a is given by (14). Hence

$$p = \hat{p}.$$

Furthermore, $A = 0$: Retailers charge no monthly fee and just pass their variable cost of wholesale power through to the consumer. The outcome of retail competition in non-linear tariffs for consumers with traditional meters is thus identical to the outcome of retail competition in linear tariffs for price insensitive customers with recording demand meters. We can thus make use of the analysis in Borenstein-Holland (2003a,b) despite the fact they they implicitly assume that all retail consumers are equipped with recording demand meters and ignore load profiling. Except in case 3, retail competition is, under load profiling, inconsistent with a Ramsey outcome.

For future reference, let U^{RC} ("RC" for "retail competition") denote the consumers' equilibrium utility:

$$U^{RC} \equiv E\left[S_i\left(D_i\left(\widehat{p}\right)\right) - \widehat{p}D_i\left(\widehat{p}\right)\right]. \tag{15}$$

Proposition 3 *Pure retail competition under load profiling delivers average wholesale power cost pricing $\widehat{p}$. The marginal price of electricity for the retail customer is therefore higher than the Ramsey price in case 1, and smaller in case 2; it is equal to the Ramsey price only in case 3.*

Remark: The Ramsey optimum can be achieved through a per customer subsidy or tax levied on retailers. Thus, let a retailer pay $\mathcal{A} + aQ$ when his customer consumes Q. The fixed charge $\mathcal{A}$ is over (or under) and beyond any

delivery, metering and customer service costs that continue to reflect responsibilities of the distribution grid owner (these costs have been normalized at zero). Faced with an access tariff $(\mathcal{A}, a)$, retailers optimally pass this tariff through to their customers ($A = \mathcal{A}$ and $p = a$). The break-even constraint of the distribution grid owner is then:

$$\mathcal{A} + E\left[(a - p_i)\, D_i(a)\right] = 0.$$

The Ramsey outcome can be obtained by setting $a = p^*$, and then $\mathcal{A}$ so as to achieve budget balance, but (except in the non-generic case 3) this requires a departure from relying on load profiled consumption to calculate the wholesale price charged to retailers, in that the variable access charge differs (except in case 3) from the consumption-weighted average pool price corresponding to the consumption induced by marginal price $p = a$.

3.3 Incumbent distributor competing with independent retailers for load-profiled customers

Finally, consider the case where the distributor is also permitted to compete for load-profiled customers. We first assume that the LSE behaves so as to maximize profits for the parent company as a whole. We then observe that nothing is altered by a ring-fencing requirement that requires the affiliate to maximize its own profits rather than those of the parent company.

a) Let us first show that the incumbent distributor's offers of the Ramsey tariff invites entry as long as $\hat{p} \neq p^*$. Suppose indeed that the distributor offer tariff (p^*, A^*). The load-profiled access charge or average wholesale power cost when the distributor serves all consumers is

$$a^* \equiv \frac{E\left[p_i D_i\left(p^*\right)\right]}{E\left[D_i\left(p^*\right)\right]}.$$

Consider an independent retailer contemplating a *small-scale entry* at some tariff $\left(\bar{p}, \bar{A}\right)$. We assume small-scale entry so that the entrant can take the access charge as given. Large-scale entry modifies the access charge that is assessed ex post, by modifying the average load profile. Alternatively, we could assume that a^* is fixed in advance based on Ramsey load profiles. This independent retailer entrant can make a positive profit provided that he offers a higher joint surplus than the Ramsey level.

Let

$$U\left(\bar{p}, a^*\right) \equiv E\left[S_i\left(D_i\left(\bar{p}\right)\right) - a^* D_i\left(\bar{p}\right)\right]$$

denote this joint surplus. Note that

$$U\left(p^*, a^*\right) = U^*.$$

Furthermore,

$$\frac{\partial U}{\partial \bar{p}}\left(\bar{p}, a^*\right) = E\left[\left(\bar{p} - a^*\right) D_i'\left(\bar{p}\right)\right],$$

and so the retailer optimally charges

$$\bar{p} = a^*.$$

The independent retailer entrant charges a higher variable price than the incumbent

$$a^* > p^*$$

if and only if

$$E\left[\left(p_i - p^*\right) D_i\left(p^*\right)\right] < 0 \quad \Longleftrightarrow \quad A^* > 0.$$

It may seem surprising that an entrant can (except in the non-generic case $a^* = p^*$ i.e., $\hat{p} = p^*$) enter against an incumbent offering the Ramsey tariff. The point is that the entrant benefits from an effective subsidy from the

incumbent, who then operates at a loss given the entry.[19] The subsidy arises as a consequence of the fact that the distributor's obligation to wholesale suppliers is equal to the aggregate metered consumption for the entire distribution system net of the load profiled consumption assigned to independent retailers.

b) Thus, assume that the incumbent distributor is regulated so as to reach the Ramsey optimum in the presence of retail competition. That is, it is instructed to maximize social welfare subject to the budget balance condition; it charges prices (p, A). The variable charge paid by retailers for each kWh consumed by their retail customers, a, is based on the average load profile of the incumbent's consumers; because the incumbent distributor can always duplicate what retailers do, we can assume without loss of generality that it serves the market (but, to serve the market, it must provide at least the net surplus offered by competitive retailers).

Let

$$V(p) \equiv U(p,p) = E\left[S_i\left(D_i(p)\right) - pD_i(p)\right]$$

with $V'(p) = -E\left[D_i(p)\right]$. The analysis in section 3.2 implies that with load profiling retailers optimally offer a linear tariff with price equal to the average wholesale power cost. And so retailers offer consumer net surplus equal to $V\left(a(p)\right)$. Note further, that because entrants prefer to offer a linear price $a(p)$ to offering marginal price p and charging a fixed fee equal to the "deficit" $[a(p) - p]\, E\left[D_i(p)\right]$,

$$V\left(a(p)\right) \geq V(p) - [a(p) - p]\, E\left[D_i(p)\right]$$

with strict inequality unless $p = a(p)$, i.e., $p = \widehat{p}$.

[19] This loss is equal to
$$E\left[(p_i - a^*)\, D_i\left(a^*\right)\right] \propto \left[E\left[p_i D_i\left(a^*\right)\right] E\left[D_i\left(p^*\right)\right] - E\left[p_i D_i\left(p^*\right)\right] E\left[D_i\left(a^*\right)\right]\right].$$

The constrained Ramsey distributor then maximizes the consumers' utility

$$\max_{\{p,A\}} \left[-A + V\left(p\right)\right]$$

subject to two constraints:

$$A + E\left[\left(p - p_i\right) D_i\left(p\right)\right] \geq 0$$

and

$$-A + V\left(p\right) \geq V\left(a\left(p\right)\right).$$

The first constraint is the distributor's zero-profit condition, and the second is the contestability constraint created by the threat of entry by pure retailers.

From the budget constraint,

$$\begin{aligned} V\left(p\right) - A &\leq V\left(p\right) + E\left[\left(p - p_i\right) D_i\left(p\right)\right] = V\left(p\right) + \left[p - a\left(p\right)\right] E\left[D_i\left(p\right)\right] \\ &\leq V\left(a(p)\right), \end{aligned}$$

with strict inequality unless $a(p) = p$, or equivalently $p = \widehat{p}$. Hence, the incumbent distributor can do no better than pure retail competition. Intuitively, the parent company by construction breaks even, and therefore the affiliate cannot do better than rival retailers, who compete with the same instruments.[20]

Remark on "ring-fencing": In the U.S. and UK there are affiliate rules that are designed to separate regulated lines of business (e.g. transmission and distribution) from unregulated lines of business (e.g. competitive generation and retailing). The rules typically require (a) cost separation to avoid cross-subsidization of unregulated lines of business by regulated lines of business,

[20]More generally, the incumbent distributor cannot deliver a net surplus to consumers in excess of $V\left(\widehat{p}\right)$ by serving some consumers but not all. To see this, note that the retail affiliate must make a non-negative profit (since a is computed so that the parent company always breaks even). By the same reasoning as above, the retail affiliate cannot offer more than $V\left(p\right) + E\left[\left(p - a\right) D_i\left(p\right)\right] < V(a)$ unless $p = a$. But if $p = a$, everyone (affiliate, independent retailers) offers retail price a, and so $a = \widehat{p}$.

(b) information transfer restrictions that limit transfers of "private information" between regulated and unregulated affiliates, (c) transfer price rules requiring any services transferred from the regulated entity to the unregulated entity to reflect either their fair market value or a regulated price and (d) equal treatment regulations that require the regulated affiliates to offer services under the same terms and condition to unaffiliated companies competing with their unregulated affiliates as they offer to their unregulated affiliates. These rules are designed to define constraints on the ability of a vertically integrated firm to maximize the joint profits of the entire enterprise. In our set-up such additional constraints on the affiliates have no impact, as the combination of break-even access charges and retail competition completely deprives the vertically integrated incumbent of any discretion.[21]

Proposition 4 *Under load profiling and retail competition, the Ramsey optimum is generically not attainable. The incumbent retailer in the constrained Ramsey optimum charges the average wholesale power cost price* $\widehat{p}$.

Remark (lagged computation of the average wholesale power cost): We have assumed that settlements occur "ex post", so a is computed on the basis of the actual aggregate consumption pattern over the period. Alternatively, one could compute a^t at date t on load profiling using date-$(t-1)$ data. Suppose that the incumbent distributor is instructed to maximize intertemporal social welfare subject to an intertemporal budget balance condition with discount factor δ, and to the contestability condition:

$$-A^t + V\left(p^t\right) \geq V\left(a\left(p^{t-1}\right)\right) \quad \text{for all } t.$$

[21] Ring-fencing in practice serves a different purpose: It aims at preventing the shifting of the costs of the unregulated affiliate company to the regulated distribution company and thus to the ratepayers.

It can be shown that the resulting constrained Ramsey price is stationary: $p^t = p^{**}$, with:

$$E\left[\left(p^{**} - p_i\right) D_i'\left(p^{**}\right)\right] = -\delta\left(\frac{\mu - 1}{\mu}\right)\left(E\left[D_i\left(p^{**}\right)\right]\right) a'\left(p^{**}\right).$$

where μ is the shadow price of the intertemporal budget balance constraint.

For $\delta = 1$, the solution is $p^{**} = \widehat{p}$ (with $\mu = \infty$). For $\delta = 0$, then $p^{**} = p^*$. And, more generally, it can be shown that the optimal policy narrows the gap between the unconstrained Ramsey price p^* and the average wholesale power cost: $p^* < p^{**} < \widehat{p}$ in case 1, $p^* > p^{**} > \widehat{p}$ in case 2, $p^* = p^{**} = \widehat{p}$ in case 3.

3.4 Incentives to install recording demand and real time meters

Let us investigate the consequences of the previous analysis for retailers' incentives to install recording demand or real time meters, starting with the Ramsey incentives. Suppose that consumers differ in the size σ of their demand: Consumer of type σ has demand $\sigma D_i(p)$ and surplus $\sigma S_i\left(D_i(p)/\sigma\right)$. There is a continuous distribution of consumers σ on $[0, \infty)$.

Consumers initially have traditional meters and cannot react to the RTP. Two types of equipments can be added to a traditional meter:

- *a recording demand meter*, costing $m > 0$, that measures and makes verifiable the consumer's RT consumption, but does not make this consumption reactive to the RTP;

- *communication* (on top of recording demand metering), costing $M > m$, that furthermore makes it possible for consumers to see and react to the RT prices through remote control of appliances and equipment.

Ramsey benchmark.

A Ramsey social planner would never install recording demand meters alone, as these do not impact behavior. The planner would equip consumers with type $\sigma \geq \sigma^*$ with communication, where

$$\sigma U^{FB} - M = \sigma^* U^* \iff \sigma^* = \frac{M}{U^{FB} - U^*}$$

(U^{FB} and U^* are given by (9) and (11)).

Retail competition.

We keep the assumption that the consumption of retail consumers with traditional meters is load profiled using the load profile of the consumers in that class. Under perfect retail competition with load profiled consumers, the consumer obtains σU^{RC} when keeping a traditional meter, $\sigma U^* - m$ when equipped with a recording demand meter, and $\sigma U^{FB} - M$ when equipped with communication.

Simple derivations yield:

Proposition 5 *(i) Under pure retail competition:*

- *Consumers with type $\sigma \geq \sigma_M^{RC}$ are equipped with communication, where $\sigma_M^{RC} < \sigma^*$ (the Ramsey level).*
- *If $\dfrac{m}{U^* - U^{RC}} \geq \dfrac{M}{U^{FB} - U^{RC}}$, then no consumer is equipped with a recording demand meter (without communication), and*
$$\sigma_M^{RC} = \frac{M}{U^{FB} - U^{RC}}.$$
- *If $\dfrac{m}{U^* - U^{RC}} < \dfrac{M}{U^{FB} - U^{RC}}$, then consumers with type σ higher than*
$$\sigma_m^{RC} = \frac{m}{U^* - U^{RC}},$$

and smaller than

$$\sigma_M^{RC} = \frac{M - m}{U^{FB} - U^*}$$

are equipped with a recording demand meter (by contrast, there is never investment in recording demand meters alone in the Ramsey benchmark). Those with type $\sigma < \sigma_m^{RC}$ *remain on a traditional meter.*

(ii) Consequently, there is more investment in meters that measure real time consumption (with and without communication) than in the Ramsey optimum. Given the inefficiencies introduced by load profiling, however investments are socially optimal.

The constrained efficiency of market-determined investment in metering equipment (part (ii) of the proposition) deserves some comment. There are really two Ramsey benchmarks, one unconstrained by retail competition and the other constrained by retail competition. If it were not for retail competition it would never be optimal to install recording demand meters if there is no price reactivity. However, the investments are socially optimal given the inefficiencies created by retail competition with load profiling.

Remark: We have assumed that with the installation of a recording demand meter comes the verifiability of the consumer's actual load curve by the local grid owner and the ISO. If this is not the case, i.e., if the retailer charges state-contingent marginal prices to the consumer, but the retailer's payment to the local grid owner is still load profiled, then there is no incentive for retailers to install meters with communication (even if M is low).[22] For, if the retailer offers price profile $\widehat{p}_\cdot$ (perhaps equal to the wholesale profile $p_\cdot$), the joint surplus of the consumer and the retailer is

$$E\left[S_i\left(D_i\left(\widehat{p}_i\right)\right) - aD_i\left(\widehat{p}_i\right)\right] \leq E\left[S_i\left(D_i\left(a\right)\right) - aD_i\left(a\right)\right].$$

[22] As suggested in Turvey (2003).

4 Price distortions: capacity obligations and ISO procurement

4.1 Price caps and capacity obligations

The existence of administrative rationing of power by system operators is sometimes used as a rationale for placing capacity obligations on LSEs. A capacity obligation requires an LSE to contract for enough capacity to meet its peak demand (plus a reserve margin in a world with uncertain equipment outages and demand fluctuations). Capacity obligations may take at least two forms. One requires LSEs to forward contract with generators to make their capacity available to the ISO during peak demand periods, leaving the price for any energy supplied by this capacity (or in a world with uncertain equipment outages and demand fluctuations the prices for operating reserves provided by this capacity as well) to be determined ex post in the spot market. Alternatively, the capacity obligations could require forward contracting for both capacity and the price of any energy (or operating reserves) supplied by that capacity during peak hours.[23]

Proposition 1 shows that rationing alone does not create a rationale for capacity obligations. Rather, there must be some reason why the spot price does not fully adjust to reflect supply and demand conditions and differs from the correct economic signal. Leaving aside procurement by the ISO for the moment, we can look in three directions. For this purpose, and like in section 2.3, we specialize the model in most of this section to *two states of*

[23]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.

nature.

Market power in the wholesale market

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

Suppose for instance that:

- baseload investment and production is competitive (as earlier),
- peakload investment and production is a monopoly.

The assumption of a monopoly in peakers is obviously unrealistic. But it is a simple way of capturing market power. Also, 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).

In the absence of price cap, a generator that has a monopoly over peak capacity would choose $K_2 = D_2(p) + \widehat{D}_2(p_2) - K_1$ so as to solve:

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

A price cap creates a shortage of peakers whenever $p^{\max} < p_2^*$, the competitive market price.[24]

To get the same level of investment and production in the second best as in the competitive equilibrium, the monopolist must receive a capacity price

[24]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.

p_K satisfying

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

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

Note that

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

and so

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

so incentives for baseload production are unchanged, *provided that off-peak plants are made eligible for capacity payments.*[25]

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

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

• *The signal for penalizing a failure to deliver is lost*: The ISO no longer has a measure of the social cost associated with a supplier's failure to deliver ($p^{\max}$ is an underestimate of this cost). Similarly, there is no objective penalty for those LSEs that underpredict their peak demand and are short of capacity obligations.[26]

• *Ex ante monopoly behavior*: If one just lets the monopolist choose the

[25] 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.

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

number of capacity contracts K_2, then the monopolist is likely to restrict the number of these contracts. Actually, 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 monopolist just exploits his monopoly power ex ante. Of course the ex ante market is more competitive than the ex post market when capacity constraints are binding.[27] How much more competitive depends on the horizon. Competition in peaking generation is likely to be intense 3 years ahead, mild 6 months ahead, and weak at $J-1$.

Two different issues have become somewhat confused in the policy discussions about capacity obligations. The first involves the nature of the contract supporting the capacity obligation. 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 the contracts supporting the capacity obligation are unlikely to be effective in mitigating market power unless the market for such contracts is more competitive than the spot market. If the capacity obligation is met with a contract that specifies both the capacity price and the energy supply price ex ante then such forward contracts can mitigate market power even if the forward market is no more competitive than the spot market.

It is well known that when generators have forward contract positions that specify the price at which they are committed to sell electricity their incentives to exercise market power in the spot market are reduced (Wolak 2000, Green 1999). For example, if a generator has contracted forward to sell all of its capacity at a fixed price p_f in each hour for the next three years it

[27] This is the view taken for example in Chao-Wilson (2003).

receives no benefit from withholding output from the spot market to drive up prices to a level greater than p_f. Indeed, in this case withholding output to drive up prices would reduce the generator's profit since it would now have to buy enough power to make up for the supplies from the capacity it withheld at an inflated price. A more controversial issue is whether and under what conditions (risk sharing considerations aside) a generator with market power in the spot market would enter into forward contracts with an overall price level lower than what they could expect to realize by not engaging in forward contracting and exercising market power in the spot market. That is, why aren't the benefits of any market power generators expect to realize in the spot market reflected in the forward contract prices they would agree to sign voluntarily as well?

Two strands of theoretical analysis have evolved to support the view that forward markets will (in essence) be more competitive than spot markets for electricity. One strand draws on papers by Allaz (1992) and Allaz and Vila (1993) that present oligopoly models in which suppliers (generators) with market power in the spot market voluntarily enter into forward contracts with prices that are lower than they would be if the suppliers only competed in the spot market. See also related work by Green (1999) and Newbery (1998). To oversimplify, the introduction of forward contracting in these models forces generators to compete both with other generators in the spot market as well as with other generators and themselves in forward markets. It creates a sort of Prisoner's Dilemma situation where individual generators voluntarily enter into forward contracts that are not in their collective interest. Chao and Wilson (2003) advance a different argument. They argue that forward markets will be more competitive (indeed "contestable") than spot markets because both incumbent generators and potential entrants

can compete in forward markets while only incumbents can compete in spot markets. See also Newbery (1998). We do not intend to resolve the issues raised by these papers here. However, from a policy perspective if voluntary forward contracting is to be relied upon to mitigate market power in the spot market, there must be some mechanism at work that does not simply allow generators to shift their market power from the spot market to the forward market.

• *A capacity payment is an insufficient instrument with more than three states of nature.* The capacity payment p_K should compensate for the revenue shortfall (relative to the socially optimal price) created by the price cap *at peak*. With many states of nature and many means of production (as in section 2.2), the capacity payment can still compensate for the expected revenue shortfall for peakers and therefore for non-peakers as well if the price cap corrects for market power at peak. However, the price cap then fails to properly correct market power just below peak. Conversely, a price cap can correct for an arbitrary number of periods/ state of nature in which there is market power, provided that the plants be dispatchable in order to qualify for capacity obligations;[28] 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

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

$i \geq i_0$ requires that:

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

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

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

Similarly, the combination of a price cap and a capacity payment cannot provide the proper signals in all states of nature to price-sensitive consumers.[29]

Remark: We have considered only aggregate uncertainty. However, a price-sensitive industrial consumer (or a an undiversified LSE) further faces *idiosyncratic* uncertainty. A potential issue then is that while the capacity payment can supply the consumer with a proper *average* incentive to consume during peak (say, when there are two aggregate states), it implies that the consumer will overconsume for low idiosyncratic demand (as she faces a "low" price p^{max} at the margin) and underconsumes in high states of idiosyncratic demand (provided that penalties for exceeding the capacity obligation are stiff). This problem can however be avoided, provided that consumers regroup to iron out idiosyncratic shocks (in a mechanism similar to that of "bubbles" in emission trading programs, or to the reserve sharing arrangements that existed prior to the restructuring of electricity systems).[30]

Proposition 6 *Capacity obligations have the potential to restore investment incentives by compensating generators ex ante for the shortfall in earnings that they will incur due to the price cap.*

[29] If there are more than three states. With three states ($i = 1, 2, 3$), the price cap can be set at p_2^*. Then $f_3\,(p_3^* - p^{max}) = p_K$ implies that $f_2\,(p_2^* - p^{max}) + f_3\,(p_3^* - p^{max}) = p_K$.

[30] 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.

(i) With at most three states of nature, the Ramsey optimum can be achieved despite the presence of market power through a combination of price cap and capacity obligations, provided that

- *off-peak plants are eligible to satisfy LSE capacity obligations and to receive capacity payments,*
- *all consumers (including price-sensitive ones) are subject to the capacity obligations,*
- *the market for peaking capacity is contestable (which probably requires capacity obligations to be imposed a few years ahead).*

(ii) With more than three states of nature, a combination of a price cap and capacity obligations is in general inconsistent with Ramsey optimality. The regulator faces a trade-off between alleviating market power off peak through a strict price cap and allowing peakers to recoup their investment; and is further unable to provide price-sensitive consumers with proper economic signals in all states of nature.

Time inconsistency / political economy

(Coming back to perfect competition), suppose that the regulator imposes an unannounced *price cap*, $p_2^{\max}$ (e.g., $p_2^{\max} = c_2$), once K_2 has been sunk. Then one would want a capacity payment to offset insufficient incentives:

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

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

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

Absence of clearing price

The third avenue is to assume a choke price: $\widehat{D}_2(p_2^*) = 0$ (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.[31]

[31] EdF (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.

4.2 Procurement by the ISO

Another potential factor leading to discrepancies between wholesale prices and social scarcity values is linked to the way system operators purchase, dispatch and charge for energy and reserves (Patton 2002). We study the implications of two such practices: out-of-merit dispatching and recovery through uplift.

4.2.1 Inefficient dispatching

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

Competitive equilibrium (indexed by a "star"):

Free entry conditions:

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

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

Supply = demand:

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

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

The competitive equilibrium is depicted in figure 1.

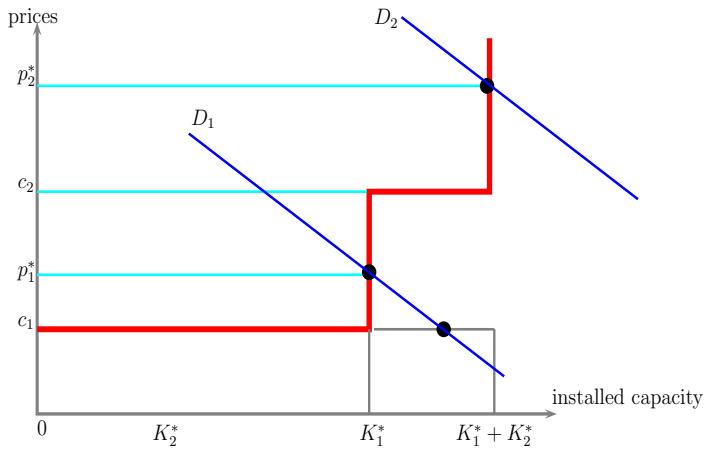


Figure 1

ISO procurement behavior

Suppose that the ISO contracts for $K_2^0 \leq K_2^*$ units of capacity and dispatches them at price 0 even off peak. This sounds strange, but more generally, as long as ISO purchases are financed externally, perverse effects arising from ISO dispatch decisions arise only if the dispatch is not economically efficient. Note also that

- we can draw D_1 such that $p_1^* = c_2$, and then some of the peak capacity must be dispatched off peak as well.

- To make things more palatable, one could imagine that state 1 is an intermediate state of demand. There would then be an off-peak state 0 with frequency f_0. As long as the off-peak price p_0^* is unaffected, one can easily generalize the analysis below.

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

Short-term impact. We analyze the short-term impact assuming a fixed capacity K_2^*. One may have in mind that K_2^0 of the K_2^* units of peaking capacity are purchased by the ISO. For given investments K_1^* and K_2^*, the short-term impact of the ISO policy is depicted in figure 2, which assumes $K_2^0 = K_2^*$:

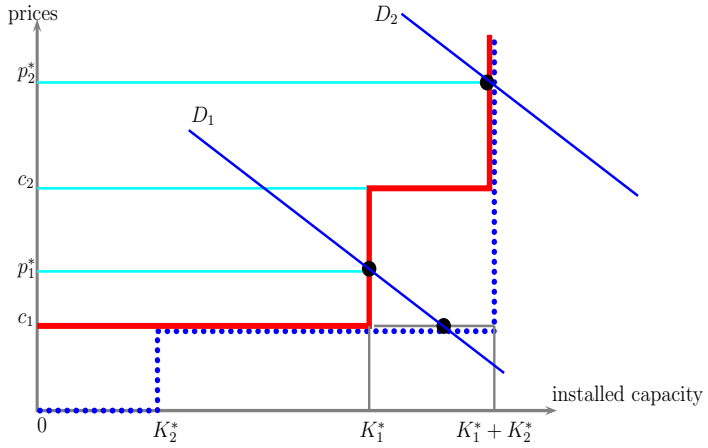


Figure 2

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

- the ISO loses

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

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

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

Proposition 7 *Suppose that ISO purchases are financed externally (i.e., not through an uplift) and are dispatched out-of-merit.*
(i) The short-term incidence of a purchase $K_2^0 \leq K_2^*$ is entirely on off-peak price and quantity: p_1 decreases, q_1 increases.
(ii) The long-term incidence of a purchase $K_2^0 \leq K_1^*$ is a substitution of off-peak units by peakers; on- and off-peak prices are unaffected.

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

Remark: When ISO purchases are financed externally, inefficiencies come solely from inefficient dispatching: Purchases introduce no inefficiency as long as they do not exceed K_2^* and the energy is dispatched only when price exceeds marginal cost.

4.2.2 Recovery through an uplift

In practice, ISO purchases are not financed through lump-sum taxation. Rather some or all of the associated costs are often at least partially recovered through an uplift. There is no general rule on how uplifts are recovered. They can be recovered monthly (often) or annually. They are typically spread across all kWh, but they can also be allocated to groups of hours (for example peak hours).

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

Suppose that the system operator purchases K_2^0 units of peaking energy forward, and dispatches the corresponding units only on peak (so that the inefficiency studied in subsection 4.2.1 does not arise). Total peaking capacity is then $K_1 + K_2$, where

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

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

The uplift t is given by

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

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

$$D_1 (p_1 + t) = K_1$$

$$f_1 (p_1 - c_1) + f_2 (p_2 - c_2) = I_1$$

$$\Longrightarrow \quad E (p) = E (p^*).$$

Peak capacity satisfies:

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

And so

$$t [K_1 + f_2 K_2] = K_2^0 I_2.$$

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

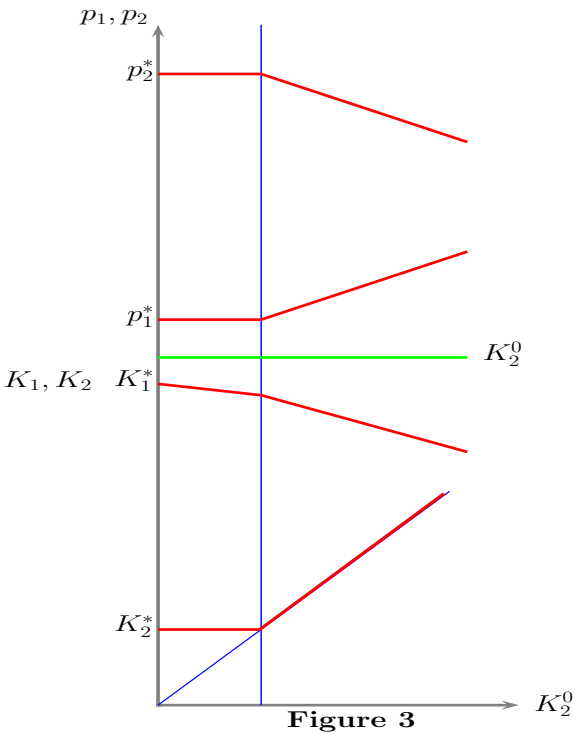


Figure 3

The results generalize to demand functions such that

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

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

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

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

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

The *off-peak* conditions are

$$D_1 (p_1) = K_1$$

and

$$f_1 (p_1 - c_1) + f_2 (p_2 - c_2) = I_1,$$

or, equivalently

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

The *peak* conditions are, as earlier:

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

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

and

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

Hence:

$$D_2 \left(p_2 + \frac{K_2^0 I_2}{f_2 (K_1 + K_2)} \right) = K_1 + K_2. \tag{16}$$

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

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

admits a single solution K and that this solution is decreasing in K_2^0.[32]

For *small purchases*, as in the case of a socialized uplift, a small purchase K_2^0 is complemented by private sector offering ($K_2 > K_2^0$) and so $p_2 = p_2^*$.

[32] One has

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

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

Given that the average price must be the same as for the free entry equilibrium, p_1 is then equal to p_1^*.

Hence, for K_2^0 small,

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

$$K_1 = K_1^*.$$

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

For *larger purchases* at some point $K_2 = K_2^0$ and private investment in peakers disappears ($f_2(p_2 - c_2) \leq I_2$). But (16) still holds. Suppose that when K_2^0 increases, p_2 increases; then p_1 decreases (as the average price must remain constant) and so K_1 increases (and so does K). For a given K, the left-hand side of (16) decreases as p_2 and K_2^0 increase. So to restore equality in (16), K must decrease, a contradiction. Hence p_2 increases.

Proposition 8 *Suppose that an uplift is levied in order to finance ISO purchases, and that the latter are dispatched in merit.*
(i) If the uplift is socialized, off-peak capacity is reduced, peak capacity may increase or decrease, and prices are unaffected for small purchases. For larger purchases, the off-peak price increases while the off-peak capacity decreases; the peak price decreases while the peaking capacity increases with the size of the purchases.
(ii) If the uplift applies solely to peak energy consumption, only peak capacity is affected (downward) for small purchases. For larger purchases, the characterization is the same as for a socialized uplift.

5 Network support services and blackouts (particularly preliminary)

This section relaxes another key assumption underlying our benchmark proposition (Proposition 1). There, we assumed that, while there may be insufficient resources and rationing, this rationing makes use of all available generation resources. This assumption is a decent approximation for, say, controlled rolling blackouts where the system operator sheds load sequentially to ensure that demand does not exceed available generating capacity. It is not for system collapses where deviations in network frequency or voltage lead to both generators and load tripping out by automatic protection equipment whose operation is triggered by physical disturbances on the network. For example, the August 14, 2003 blackout in the Eastern United States and Ontario led to the loss of power to over 50 million consumers as the networks in New York, Ontario, Northern Ohio, Michigan and portions of other states collapsed. Over 60,000 MW of generating capacity was knocked out of service in a few minutes time. Most of the generating capacity under the control of the New York ISO tripped out despite the fact that there was a surplus of generating capacity to meet demand within the New York ISO's control area. Full restoration of service took up to 48 hours. (U.S.-Canada Power System Outage Task Force, 2003). The September 28, 2003 blackout in Italy led to a loss of power across the entire country and suddenly knocked out over 20,000 MW of generating capacity. Restoration of power supplies to consumers was completed about 20 hours after the blackout began (UCTE, 2003).

Conceptually, there is a key difference between rolling blackouts in which the system operator sequentially sheds relatively small fractions of total demand to match available supplies in a controlled fashion and a total system

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

It is useful here to relate this economic argument to standard engineering considerations concerning operating reserves (OpRes) and ICAPs. In addition to dispatching generators to supply energy to match demand, system operators schedule additional generating capacity to provide operating reserves (OpRes). Operating reserves typically consist of "spinning reserves" which can be fully ramped up to supply a specified rate of electric energy production in less than 10 minutes and "non-spinning reserves" which can be fully ramped up to supply energy in up to 30 minutes (60 minutes in some places). Operating reserves are used to respond to sudden outages of generating plants or transmission lines that are providing supplies of energy to meet demand in real time sufficiently quickly to maintain the frequency, voltage and stability parameters of the network within acceptable ranges. Additional generation is also scheduled to provide continuous frequency regulation (or automatic generation control) to stabilize network frequency in response to small instantaneous variations in demand and generation. These ancillary network support services require scheduling additional generating capacity equal to roughly 10-12% of electricity demand at any point in time. In the U.S., regional reliability councils specify the requirements for frequency regulation and operating reserves, as well as other ancillary services such as

reactive power supplies and blackstart capabilities, that system operators are expected to maintain. Pending U.S. legislation would make these and other reliability standards mandatory for system operators.

Installed capacity or "ICAP" obligations are in place in New England, New York and PJM. They have been proposed for California and were included as a proposal in FERC's Standard Market Decision rulemaking. ICAP obligations require load serving entities to have forward contracts for enough generating capacity to meet their forecast peak demands plus a reserve component (e.g 118% of forecast peak demand). Typically, these ICAP obligations require LSEs to have forward contracts for capacity (the capability to supply energy to the network), but whether or not this is accompanied by a forward contract on the price of the energy to be supplied by this capacity in real time is up to the LSE. Capacity that has been identified by an LSE to meet its ICAP obligations must be made available to the ISO when the ISO calls for it. ICAP obligations are enforced with deficiency penalties and moral suasion. LSEs may use ICAP capacity that they have under contract to meet their OpRes obligations in real time as well.

Let us use a simple model of OpRes in order to analyze the various issues at stake. To keep modeling details to a minimum, 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 IK and marginal cost c.

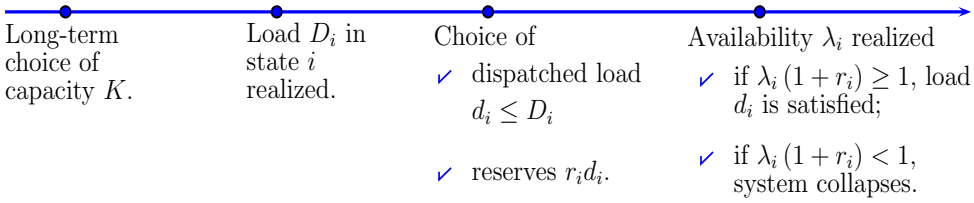
The key innovation relative to the benchmark model is that the extent of scarcity is not fully known at the dispatching time. We formalize this uncertainty as an uncertain *availability* factor $\lambda \in [0,1]$. That is, a fraction $1-\lambda$ of the capacity K will break down. The distribution $H_i(\lambda)$(with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

and $H_i(1) = 1$) can be state-contingent.[33] There may be an atom in the distribution at $\lambda = 1$ (full availability), but the distribution has otherwise a smooth density $h_i(\lambda)$.[34] 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)

The timing goes as follows:



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. We assume that mere availability 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). If

$$\lambda_i\,[(1 + r_i)\,d_i] < d_i,$$

the system collapses, and no energy is produced or consumed.

a) *Social optimum*

[33] 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.

[34] 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.

A Ramsey social planner would solve:

$$\max_{\{K,d_\cdot,r_\cdot\}} \left\{ E\left[\left[1 - H_i\left(\frac{1}{1+r_i} \right) \right] (\mathrm{v} - c) - s\,(1+r_i) \right] d_i - KI \right\}$$

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

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

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

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:

$$\frac{h_i}{(1+r_i)^2}(\mathrm{v} - c) - s = \nu_i, \tag{17}$$

$$[1 - H_i]\,(\mathrm{v} - c) - s\,(1+r_i) = \mu_i + (1+r_i)\,\nu_i, \tag{18}$$

and

$$E\,[\nu_i] = I. \tag{19}$$

Specializing the model to the case in which H_i is state-independent,[35] let us first analyze the optimal dispatching, as described by (17), (18) and (19).

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

The Ramsey optimum is depicted in Figure 4.

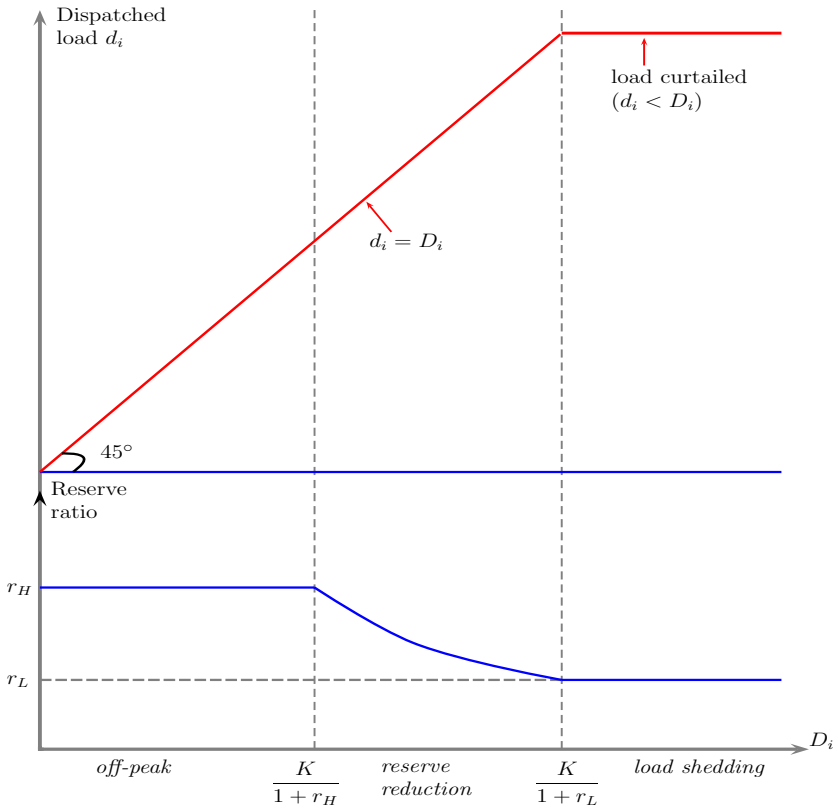


Figure 4

Off-peak (D_i small), there is excess capacity and $\nu_i = 0$. Hence

$$r = r_H$$

where

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

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

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

The off-peak region is defined by:

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

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

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

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

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

which from our assumptions has a unique solution:

$$r_L < r_H.$$

The optimal investment policy is then given by:

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

b) *Implementation*

First, note that reserves are a public good. Network users take its reliability as exogenous to their own policy and thus are unwilling to contribute to reserves. The market-determined level of reliability is therefore the size of the atom of the $H(\cdot)$ distribution at $\lambda = 1$. Thus, the market solution leads to an insufficient level of reliability.

In order to obtain a proper level of reliability, the system operator must force consumers (or their LSE) to purchase a fraction r_i of reserves for each unit of load. There can then be two markets, one for energy at price p_i and one for reserves at price p_i^R (related through the arbitrage condition $p_i^R = p_i - c$ (Stoft 2002) in the absence of collapses).

Does this market mechanism cum regulation of reserve ratios generate enough quasi-rents to induce the optimal investment policy? In the following, we will normalize $c = 0$, so as to avoid discussing the accounting of extra marginal costs incurred when reserves are dispatched and marginal cost savings when the system collapses. Off-peak ($D_i < K/(1 + r_H)$), the price of energy is 0, and there are no quasi-rent.

When load is curtailed ($D_i > K/(1 + r_L)$), then the market price of energy is $\mathrm{v}/(1 + r_L)$ (since consumers are willing to pay v per unit, but must buy 1 unit in the energy market and r_L units in the OpRes market to obtain 1 unit if the system does not collapse). Thus, generators obtain, as they should, quasi-rent:

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

in this region.

The intermediate region is more complex to implement through an auction-type mechanism. In the absence of price-responsive load, the supply curve and the total demand curve (energy plus reserves) are vertical and identical.

Hence a small mistake in the choice of reserve ratio creates wild swings in the market price (from 0 to $\mathrm{v}/(1+r_i)$). In particular, the system operator can bring price down to marginal cost without hardly affecting reliability.

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

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

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

to generators in this region. This regulation too involves substantial discretion, however.

6 The joint interruptibility problem (particularly preliminary)

We last discuss the remaining key assumption underlying the benchmark proposition: that different users can choose different levels of priority in rationing. For this to be doable, the system operator must be able to shut off consumers individually.

There is no theoretical reason why individual customers cannot be rationed. It requires installing communications and control equipment between the customer's connection to the network and the control center. However, this equipment is costly. As a practical matter, except for very large customers that have direct control equipment, most directed interruptions must occur at points on the network ("zones") that can be controlled by the distribution network operator.[36] The affected zone has (a) customers served by multiple LSEs that compete with one another (so every house on a street can be "served" by a different LSE) and (b) customers with heterogeneous preferences.

An optimal dispatch when zones but not individual consumers are controlled by the system operator must elicit each zone's *aggregate* willingness to pay for being served. From the point of view of the set of LSEs and industrial users in a given zone, reliability is a *public* good.

In principle, one can make use of the theory of public goods in order to design incentive-compatible mechanisms of elicitation of individual preferences

[36]In reality, system operators generally try to squeeze out all of the price sensitive demand first before they start rolling blackouts. This may not be optimal of course. There is also some priority rationing in that circuits with hospitals and fire stations, etc. will often be placed on a "do not blackout list." In this case, all customers on the same circuit get the benefit of being near a fire station or hospital. This example illustrates the fact that different consumers may have different values of lost load, and that furthermore the dispatcher cannot fine-tune the intensity of rationing.

for reliability.[37] For instance, one could use the Clarke-Groves scheme.[38]

Besides transaction costs, there is under retail competition a major snag with such zonal voting mechanisms. Competing retailers' profit in a given zone depends only on their *relative* quality of their offer as compared with their competitors'. A retailer that bids for reliability increases the quality of service to its retail consumers, but it also increases its rivals' quality of service by the same amount, bringing no extra profit.[39] This problem does not arise among a monopoly distributor and non-competing industrial users.

We thus conclude that *the joint interruptibility problem is particularly difficult to solve through market mechanisms when retail competition is allowed.*

---

[37] See Green-Laffont (1979a,b) for the general theory of public goods.

[38] Suppose that, due to a shortage in supply, the ISO must shut down one of cities A,B,C,... To simplify computations, cities demand the same load. Within city A, say, there are n users, each demanding 1 unit of load and having valuations (VOLL) v_i, which are private information. These users can either be price-sensitive, industrial users or LSEs serving price-insensitive users. Let the ISO shut down the city with the lowest total declared willingness to pay. That is, city A is served if and only if

$$\hat{V}_A \equiv \sum_{i \in A} \hat{\mathrm{v}}_i \geq \hat{V}$$

where $\hat{V}$ is the lowest total declared willingness to pay among other cities. City A then pays $\hat{V}$. The problem then boils down to a standard public good problem (the cost of getting the public good is $\hat{V}$-possibly unknown to members of city A, but this does not matter).

In particular, use can be made of Clarke-Groves mechanisms : Member i of city i pays

$$\begin{cases} \hat{V} - \sum_{j \neq i} \hat{\mathrm{v}}_i & \text{if} \quad \hat{\mathrm{v}}_i + \sum_{j \neq i} \hat{\mathrm{v}}_j \geq \hat{V} \\ \\ 0 & \text{otherwise.} \end{cases}$$

Telling the truth ($\hat{\mathrm{v}}_i = \mathrm{v}_i$) is then a dominant strategy. [The Clark-Groves mechanism does not balance the ISO's budget, but a variant of it (the d'Aspremont-Gerard Varet scheme) does so in expectation.]

[39] This is best seen when considering the following timing: First, LSEs bid for reliability $\left(\alpha_i^k\right)$ in zone k. Second, given the resulting $\left\{\alpha_\cdot^k\right\}$, they compete for retail consumers in zone k as in sections 3 and 2. Given that they make no profit at stage 2, LSEs aim at mainimizing expenditure at state 1 (they have de facto willingness to pay $\hat{\mathrm{v}}_i = 0$ in reference to the previous footnote).

References

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

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

[3] Borenstein, S., and S. Holland (2003a) "Investment Efficiency in Competitive Electricity Markets With and Without Time-Varying Retail Price," CSEM WP 106R.

[4] Borenstein, S., and S. Holland (2003b) "On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices," CSEM WP 116.

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

[6] Chao, H. P., and R. Wilson (2003) "Resource Adequacy and Market Power Mitigation via Option Contracts," mimeo, Stanford University.

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

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

[9] Green, J., and J.J. Laffont (1979a) *Incentives in Public Decision Making*, Amsterdam: North Holland.

[10] —— , eds. (1979b) *Aggregation and Revelation of Preferences*, North-Holland.

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

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

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

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

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

[16] —— (2003) "The Demand for Operating Reserves: Key to Price Spikes and Investment," *IEEE*.

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

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

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

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

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

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

Appendix: Generalization to consumer heterogeneity (incomplete)

For expositional simplicity, we have assumed that consumers are homogenous (perhaps up to a size factor σ). This appendix briefly investigates the implications of consumer heterogeneity for retail competition (section 3).

Suppose that there are different classes of consumers $j \in [0,1]$ with state-contingent demands $D_i^j(p)$ and state-contingent surplus $S_i^j\left(D_i^j(p)\right)$. Let n^j denote the frequencies of consumers of type j, and $E_j[\cdot]$ denote the expectations (the expectations with respect to the state of nature are now labeled $E_i[\cdot]$).

With non-reactive consumers, the *Ramsey optimum* maximizes the sum of the net surpluses:

$$\max E_i\left[E_j\left[S_i^j\left(D_i^j\left(p^j\right)\right) - p_i D_i^j\left(p^j\right)\right]\right]$$

yielding for each j

$$E_i\left[\left(p^j - p_i\right) D_i^{j\prime}\left(p^j\right)\right] = 0. \tag{3'}$$

Let $Q^j \equiv E_i\left[D_i^j\left(p^j\right)\right]$ denote type j' s total consumption in this Ramsey optimum, and assume that this consumption increases with j. Let $\boldsymbol{S}^j(Q)$ be defined by:

$$\boldsymbol{S}^j(Q) = \max E_i\left[S_i^j(D_i)\right] \text{ subject to } E_i[D_i] \leq Q.$$

So $\boldsymbol{S}^j\left(Q^j\right)$ denotes the Ramsey average gross surplus of consumer j. Provided that the following sorting condition holds:

$$\frac{\partial}{\partial j}\left(\frac{\partial \boldsymbol{S}^j}{\partial Q}\right) > 0,$$

The Ramsey allocation is implementable. The tariff $T(Q)$ that implements it, however, need not be equivalent to a menu of two-part tariffs.[40]

The treatment of *pure retail competition* with traditional meters (load profiling) by contrast follows the lines of section 2.2. Retailers charge $p = a$ to their consumers, and the equilibrium price $\hat{p}$ is still given by (13) by setting

$$D_i(\hat{p}) \equiv E_j\left[D_i^j(\hat{p})\right].$$

[40]Here is a counterexample: Suppose that $D_i^j = k + i + j - p$. Then the Ramsey price is $p^j \equiv E_i[p_i]$ and independent of j. Q^j increases with j and the sorting condition holds. However, given that p^j is the same for all j, incentive compatibility requires that the fixed fee A^j be also the same for all j. It may not be feasible to simultaneously attract the low types ($j = 0$) and to cover the deficit $E_j\left[E_i\left[\left(p^j - p_i\right)D_i^j\right]\right]$ on purchases in the wholesale market.