Flexibility in Power System: Market Design Matters

Dongchen He[*], Bert Willems[†]

October 29, 2025

Latest Version Here

Abstract

The growing share of renewable energy requires sufficient investment in power system flexibility. In this paper, we frame a three-stage peak-load pricing model consisting of investment, commitment, and production, considering that electricity generation is costly to adjust on short notice. The results demonstrate the importance of increasing time granularity in electricity markets with efficient state-contingent prices. Adapting the idea of real options theory that waiting is valuable, flexible firms avoid producing in the low-demand state and earn a premium to recoup investment costs.

On top of that, this paper discusses the efficiency of alternative market designs in the investment of flexible assets. In the absence of an efficient real-time market, the forward price results in a distortion in technology investment. This distortion, in theory, can be corrected by an options market, while any centralized auction fails to achieve the optimum. Finally, this work briefly illustrates the effect of demand flexibility, showing that an increase in demand response does not necessarily reduce the reliance on production flexibility if rationing is done randomly.

Keywords: Flexibility, Real-Time Prices, Reserves, Uncertainty, Electricity, Market Design

JEL: D82, L23, L94

[*]He is at the School of Economics and Management, Tilburg University.

[†]Willems is at the School of Economics and Management, Tilburg University, Université catholique de Louvain and Toulouse School of Economics.

1 Introduction

This paper analyzes the value of generation flexibility and how market design affects investment decisions when quick adjustment is expensive. This question is particularly pertinent to electricity markets, demand of which is so unpredictable, while continuously balancing the market is vital to social and economic development; the energy transition further raises the concern about the reliability of the electricity supply. According to the report "Net Zero by 2050" from the International Energy Agency (hereafter: IEA), renewables are expected to generate 88% of global electricity in 2050. Solar PV and wind make up nearly 70% of the total share. As a reference, 29% of electricity is supplied by renewable sources in 2020, and about two-thirds is from hydro-power.

With an ambitious investment in global intermittent renewable energy sources (IRES) over the next three decades, IEA predicts a fourfold increase in the demand for flexibility.[1] Enhancement of flexibility in power sector includes supply-side flexibility such as retrofitting existing thermal sources and building new flexible plants, as well as demand-side flexibility, storage[2], grid reinforcement, etc. Along with the technology changes, new designs for electricity markets, especially the real-time market and ancillary markets, are required to incentivize operations and investment of flexible units.

IEA defines flexibility as "*the ability of a power system to reliably and cost-effectively manage the variability and uncertainty of demand and supply across all relevant timescales*". Hence, besides the physical characteristics of shorter start-up times and higher ramp rates, it is equally important to economically run flexible units. Flexibility in this context refers to an economic terminology: adjustment cost. Therefore, this paper establishes a model based on adjustment cost and assumes that technologies that cannot adjust on short notice have an infinitely large adjustment cost. Conventional generators are not equally flexible, ranging from nuclear plants, which are inflexible, to coal plants, oil-fired plants, and gas turbines that are rather flexible.

[1] Flexibility requires more than double by 2030 and 7 times by 2050 in the EU, according to the European Commission. See https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/future-eu-power-systems-renewables-integration-require-7-times-larger-flexibility-2023-06-26_en.

[2] We do not explicitly model storage in this paper, but as a way of providing flexibility, the investment decision in storage is similar to other fossil flexibility and the main result in our paper still applies. To store energy is equivalent to purchasing power, and to release it is a way of production. The difference, however, is that storage has to switch the role of power producer and buyer, so it cannot make an independent decision for each state.

Retrofits or investments in new flexible units incur a higher investment cost in exchange for reducing adjustment costs. For example, the hydrogen-powered system is very flexible, but expensive. Solar panels with batteries are flexible, but more costly than standard ones without a battery. Coal-fired power plants can improve their operational flexibility by retrofitting with a steam turbine and thermal energy storage. Therefore, flexibility is cheap in adjustment, while expensive in investment and/or production. An optimal technology mix is to provide sufficient flexibility cost-efficiently, the key to which is to properly price flexibility.

Flexibility is not a new topic in power markets. A batch of engineering-oriented studies evaluate different approaches, such as storage, demand side management, and grid expansion, to provide flexibility with higher penetration of intermittent renewables (Lund et al., 2015; Kondziella and Bruckner, 2016; Denholm and Hand, 2011). Brijs et al. (2017) assess instruments like price floor or cap in short-term markets and their impacts on the supply of flexibility, providing simple numerical results. With an abundance of economic literature (Lucas, 1967; Gould, 1968; Schramm, 1970; Ito and Reguant, 2016; Hortacsu and Puller, 2008) discussing the impact of inflexibility on production and/or investment, few look into the effects of market design. Furthermore, real-time or balancing markets receive less attention at a time when renewable sources are negligible. Reserve markets are designed to induce short-term flexibility, but efficiency in the long term is not well addressed. This work aims to fill this gap to understand the effects of design on both real-time flexibility supply and long-term investment by building a stylized theory model.

Power prices are usually determined before demand is known for the sake of coordination between power suppliers and operators, and risk hedging. However, this paper, by extending the traditional two-stage peak-load pricing model to a three-stage game consisting of investment, scheduling (commitment), and production, shows that payment certainty would impede investment in flexibility because they are not able to earn more than inflexible ones. Van Der Weijde and Hobbs (2012) quantify the multi-stage settlement and validate uncertainty in transmission planning. The model in Anupindi and Jiang (2008) has a similar timing as our work, but they concentrate on the strategic equilibria in a duopoly competition where firms are both flexible or inflexible. Nevertheless, our model proposes a monopolistically competitive scenario and reveals the impacts of market organization. We address two main research questions: *(1)What is the efficient pricing and investment of flexibility? (2) Which market designs can achieve this efficiency?*

The results show that peak-load pricing can be applied to a competitive real-time market to correctly reflect the cost and value of flexibility. In the absence

of real-time pricing, the forward price distorts the investment decision. Complementing with a reserve market in which the system operator (SO) provides a menu of contracts can restore the optimum.[3] Nevertheless, the existing reserve market design, such as a forward integrated auction of energy and reserves (Oren and Sioshansi, 2005; Ehsani et al., 2009), non-linear scoring auction, or uniform pricing for reserves (Chao and Wilson, 2002) leads to flexibility investment distortion.

Our contribution is two-fold. First, we develop a variation of the peak-load pricing model to price the value of flexibility. We model an extra stage for commitment, so production under loss in the short term is allowed. Compared to the traditional two-stage model, our supply function is steeper, and the price distribution is much more volatile. Hortascu and Puller (2008) show that supply curves of small generators are rather inelastic in the real-time market. Ito and Reguant (2016) find evidence that for fringe suppliers, adjustment in intra-day markets costs 5 to 10 times more than forward scheduling. We build a theoretical framework to incorporate those observations, and we believe this is necessary when more uncertainty is introduced in the energy transition.

Second, as a key insight of this paper, we show that firms that sell reserves should earn a technology-specific flexibility premium on top of the opportunity cost of not trading in the forward energy market and the related adjustment costs. The flexibility value is called option premium or value of waiting in real options theory (Trigeorgis, 1996; Schwartz and Trigeorgis, 2004), defined as the difference between real options value and net present value.[4] Reserves are not only a backup to unanticipated supply or demand shocks, but can be strategically operated: firms trade off the flexibility premium earned from providing reserves and adjustment costs saved by serving energy. This finding implies the failure of current reserve markets to properly incorporate this premium, which would lead to insufficient investment in flexibility.

Note that the social optimum is hardly obtained by a combination of market-based forward and reserve markets. Hence, this paper cautiously supports a real-time electricity market in which flexible firms self-schedule and adjust their production to price signals instead of explicitly selling reserves in a reserve market. In other words, in the absence of risk aversion and market power, when efficient real-time pricing is available, a separate reserve market is unnecessary. [5]

[3] In the context of this paper, a reserve market plays the role of an option market.

[4] In this paper, those terms are used interchangeably.

[5] As a real-time market is not fully efficient in practice and the system operator is averse to power outage, a common reliability criterion requires the system operator to at least reserve capacity that is able to keep the grid stable in the event of an unexpected outage of the largest

In practice, it is difficult for all transactions to be settled nearly instantaneously. When approaching real-time, things become more certain, so a forward market and an intra-day market are rationalized to trade power that is likely to be consumed, and only the most flexible generations are reserved and traded in real-time, which is called a balancing market.[6] The point is, less flexible firms can trade earlier, and flexible firms need a real-time price signal to induce them to trade near production. Flexible technology makes profits from uncertainty, and the strategy to sell reserves beforehand to address uncertainty could backfire unless an elaborate payment scheme is designed. Moreover, we simplify risk preference and market structure in order to disentangle the effects of market design on flexibility, which asserts the significance of real-time market to compensate for flexibility, but there is no way to repudiate the forward market as a way to hedge risk and mitigate market power.

In the end, this paper exploits the relation between demand response and supply flexibility. One consensus that production flexibility is so important is the lack of demand-side management. When consumers cannot react to prices, firms have to adjust supply to balance the market. Otherwise, rationing happens. Hence, it is intuitive to reckon that larger demand flexibility would decrease the investment in flexible capacity. However, this paper states that demand and supply flexibility are not always substitutes. If rationing is random over consumers, the increase in demand response does not necessarily reduce the need for production flexibility.

The rest of the paper is organized as follows. Section 2 provides an overview of the relevant literature. Section 3 describes the model. Section 4 illustrates the equilibrium of different market designs, and section 5 concludes.

2 Literature Review

This work relates to three strands of literature: peak-load pricing model, adjustment costs and inflexibility, as well as reserve markets. In this section, we discuss the development of literature and highlight the contribution of this paper.

generator. However, as proven in this paper, a short-term reserve auction does not sufficiently reimburse for reserve provision.

[6]Another proposal that is adopted by Texas and Ontario is to require suppliers to submit their commitment or prediction day ahead, but clear in real-time.

2.1 Peak-load Pricing

Peak-load pricing model was developed from 1949 to determine the efficient pricing and investment under demand variation. Investment costs are irreversible, and some capacity is only utilized during peak time.

Modeling periodic and deterministic demand supplied by a single technology, Boiteux (1960)[7]shows the off-peak price is equal to marginal production cost and consumers during peak hours pay for both production and capacity costs. Crew and Kleindorfer (1976) extend this model to multiple technologies, and the new insight is that both off-peak and peak consumers should pay for capacity on top of production costs. Joskow and Tirole (2007) follow this setup and extend it to a continuum of technologies. It is also recognized that demand is not only periodic, but also uncertain (Visscher, 1973; Carlton, 1977; Brown & Johnson, 1969; Crew & Kleindorfer, 1976). The problem becomes choosing capacity and price before demand is realized. The results highly depend on the model setup.

Visscher (1973) shows that with random rationing, optimal pricing and investment would result in a price lower than long-run marginal cost (hence, allowance is needed to guarantee the zero-profit condition), but capacity is the same as efficient rationing. Carlton (1977) proves that pricing depends on the way uncertainty enters the demand function. If uncertainty enters the demand curve in an additive term, price is lower than long-run marginal cost, but the conclusion is reversed with multiplicative demand uncertainty.

The theoretical progress has stopped since then. With uncertainty, the equilibrium mentioned above is sub-optimal since it lacks state-contingent prices and output. Two points are missing in the literature. First, it assumes all technologies are flexible, while in reality, real-time adjustment is rarely cost-free, and ignoring this cost would lead to inefficient investment. Second, even though technologies are fully flexible, the literature fails to reflect the flexibility value in price. It is important to realize that firms trade off the flexibility of postponing production decisions and the value of commitment by saving adjustment costs, which in turn affects capacity decisions.

This paper extends the classic two-stage peak-load pricing model to three stages, of which production commitment is between investment and production, arguing that efficient real-time prices should reflect the cost of adjustment and the value of flexibility. We model both periodic and uncertain demand net of renew-

[7]This article was translated by H. W. Izzard from an article in French which appeared in the *Revue générale de l'électricité* in August, 1949.

ables, but the focus of this paper is efficient investment in flexibility rather than investment at peak time.[8]

2.2 Adjustment Costs

This paper also connects with the long-standing literature on adjustment costs. Adjustment costs refer to costs incurred when a decision is changed, accounting for slower changes in inputs in response to external shocks. This concept is also widely used in analyzing stocks (Hay, 1970), capital investment (Lucas, 1967; Gould, 1968; Schramm, 1970), and labor demand (Jaramillo et.al., 1993).

Lucas (1967) clarifies that there are both fixed and variable inputs, so long-run and short-run supply behavior are distinct. Fixed inputs cannot be changed in the short term, and adjustments to demand are staggered. Therefore, the long-run equilibrium is not the minimum point of a U-shaped cost function, but it also includes the costs of approaching and keeping it.

Adjustment costs are analyzed in econometric studies of labor demand, and the core discussion is the structure of costs: whether hiring or firing costs are symmetric? While the standard assumption of adjustment cost is a symmetric and quadratic function, data collected from Italy (Schramm, 1970), the Netherlands, and the UK (Pfann & Verspagen, 1989; Pfann & Palm, 1993) reject this hypothesis. There is also a strand of energy economics literature considering inflexibility in the real-time power market by introducing a steeper supply curve in real-time compared to a forward market (Ito & Reguant, 2016; Hortaçsu & Puller, 2008).

The structure and level of adjustment costs should concern economists of many stripes. To be able to predict the effect of shocks, economists should know both the source of adjustment costs and how they are reflected in equilibrium behavior. Dispatchable generators are an indicator of supply-side electricity flexibility, and economists should ensure enough flexibility while controlling the adjustment costs to an acceptable level. This paper assumes (a)symmetric adjustment cost and enriches the model by adding (a)symmetric investment costs in flexibility. That is, not only does deviation incur adjustment cost, but firms have to invest in flexibility to be able to deviate.

[8]Peak-load technologies are often used to provide flexibility, but this does not mean they are equivalent. Peak-load technologies increase supply for anticipated peak demand, while flexibility expands or curtails output to sudden demand and supply change.

2.3 Reserve Markets

In electricity markets, power supply adjustment is largely done by reserve, one of the ancillary services centrally procured to satisfy demand when supply and demand uncertainty that would otherwise lead to a blackout (Cramton, 2017). Hence, reserves are usually regarded as a way to provide reliability (Bushnell & Oren, 1994; Cramton, 2017; Wilson, 2002), especially when renewable energy is integrated into the electrical grid (Sedzro et al., 2018).

Joskow and Tirole (2007) consider operating reserves a public good, so they should be procured centrally to prevent a full system breakdown. In their model, the optimal dispatched load is known, but a fraction of capacity may unexpectedly fail in real time. Hence, it is necessary to provide reserves to avoid system collapse. Reserves play a role in providing additional capacity and avoiding a possible huge loss, which is a natural and standard reliability consideration.

On top of that, reserves are also viewed as a financial hedge [9] to deal with spot price uncertainty. In two related papers (Kleindorfer & Wu, 2005; Anderson et al., 2017), reserves are used as an options contract. Instead of a "backup" role, reserves are strategically substitutes for energy in the real-time market, and market equilibrium will result in an optimal allocation between reserves and energy to maximize utility, based on real-time price distribution. Hence, if the real-time price is very low, power buyers do not reserve any capacity and only rely on the spot market. Moreover, they assume there are multiple strategic suppliers in a reserve market and more nonstrategic suppliers in the spot market and solve the bidding strategies of reserve suppliers, which mirrors the results in Chao and Wilson (2002).

This paper proposes another reason to deploy reserves: it is a way to provide and price flexibility in the absence of a real-time market. Most research focuses on the auction design of reserve markets, analyzing the bidding mechanism and bidder strategies that result in short-term efficient operations such that energy is dispatch in a merit order[10]and as less as possible information rent extraction. The difficulties are to reimburse both the capacity and production parts of reserves and to deal with asymmetrical information on costs. The main finding from Chao and Wilson (2002) is that capacity and production should be bid separately. The strategic bidding only uses the capacity part, and energy supplies are paid the

[9]For instance: the Short Term Operating Reserve (STOR) in the UK.

[10]The merit order is a way of ranking dispatch of electricity, based on ascending order of price which should reflect the order of their short-run marginal costs of production and sometimes pollution(and other externality), together with amount of energy that will be generated.

spot price. A nonlinear scoring rule with discriminatory pricing also works if generators agree with the system operator on the probability distribution of energy calls (Bushnell & Oren, 1994). Oren and Sioshansi (2005) propose an integrated market for both energy and reserves in which the activated reserves are paid the same price as energy, and reserves not activated receive a capacity payment equal to the difference between the clearing price and their own bid. This design is simple and incentive compatible, but it does not consider any direct cost of keeping capacity, and the only economic cost is the opportunity cost of not being sold as energy.

This paper argues that if adjustment costs can be properly reflected in the real-time price, a separate reserve market is not necessary anymore. A complicated reserve market design can be replaced by a good signal: real-time price. However, if real-time pricing is not possible, it is crucial to include the technology-specific flexibility value in the reserve payment. This finding shows that none of the uniform pricing auctions proposed by Chao and Wilson (2002), the integrated auction by Oren and Sioshansi (2005), and the nonlinear scoring auction by Bushnell and Oren (1994) can guarantee a proper level of flexibility.

3 Model

We develop a three-stage model of electricity generation under uncertainty. In the first stage (investment), the generators choose how much capacity to install from a range of technologies. In the second stage (unit commitment), they commit to production quantities and reserve levels, i.e., the amount of operational flexibility they schedule, based on partially revealed information about future demand. In the third stage (real-time operation), the actual demand is realized. Generators adjust their output within the limits of their previous commitments. Generation technologies vary in capital cost, marginal production cost, and flexibility costs associated with scheduling reserves and adjusting production in real time.

Electricity demand is uncertain and changes between stages of the model. Before committing to production, generators receive some information about future conditions, such as time of day, demand, and weather forecasts, but do not observe the exact demand realization. In the unit commitment stage, generators manage this uncertainty by scheduling production and upward and downward reserves. These reserves allow them to adjust the output upward or downward once actual demand is realized.[11] Overcommitting production may result in costly downward

[11] In practice, reserves are not only characterized by direction (upward and downward), but

adjustments, while undercommitting may require activating expensive upward reserves. In the investment stage, generators can manage uncertainty by investing in technologies that have lower flexibility costs. However, more flexible technologies may involve higher capital or operational costs.

On the demand side, there are two categories of representative consumers: price-sensitive and price-insensitive consumers. Price-sensitive consumers observe real-time market conditions and can adjust their level of consumption accordingly. Price-insensitive consumers do not respond to real-time prices, although their consumption may be rationed when market conditions require it. Price-sensitive consumers buy directly in the wholesale market. Price-insensitive consumers are served by competitive retailers. Retailers do not incur costs for retail activities. They buy electricity in the wholesale market and sell it to consumers under two-part tariffs, corresponding to a marginal energy fee and a lump sum fee.

In our model, inflexibility arises from both the supply side and the demand side. Supply inflexibility reflects the cost of adjusting output on short notice. On the demand side, price-insensitive consumers do not adjust consumption in response to real-time prices. Their demand is perfectly inelastic, unless consumers are rationed.

We assume that all parties in the model are risk-neutral. Generators, retailers, and consumers maximize expected profit and expected net utility. Hence, there is no need for contracts to reallocate risks across agents.

3.1 Timing and Uncertainty

We now introduce more formally the timing and uncertainty structure of the model. See Figure 1. We represent the gradual resolution of uncertainty by a two-dimensional vector $\xi = (\omega, \varepsilon)$ which describes a possible information trajectory. The interim state ω represents the information that is available at the unit-commitment stage ($T = 2$). The variable ε denotes the final realization of demand at the real-time stage ($T = 3$). The joint density function of information trajectories is given by $h(\xi)$, defined as

$$h(\xi) = g(\omega) \cdot \phi(\varepsilon|\omega), \tag{1}$$

also by response time and the duration of activation: Primary reserves (Frequency Containment Reserves) act within 10 seconds, Secondary reserves (Automatic Frequency Restoration) are activated within 30 seconds and may last 5 to 15 minutes and Tertiary reserves (Manual Frequency Restoration) are dispatched within 5 to 15 minutes and may last up to a few hours

where $g(\omega)$ is the probability density of interim states, and $\phi(\varepsilon|\omega)$ the conditional distribution of the final demand realization ε given ω.

When making investment decisions, firms maximize expected profit over all possible information trajectories, as represented by $h(\xi)$. This is analogous to real option valuation methods in finance, where Monte Carlo simulations are based on simulated state trajectories. The value of flexibility at the interim state ω depends on the dispersion in $\phi(\varepsilon|\omega)$. Commitment decisions are made conditional on the interim information ω, while real-time actions depend on the full information trajectory ξ.

In our model, the interim state ω captures both predictable variation, such as hour-of-day and day-of-week effects, as well as forecasted information about demand, including weather conditions.

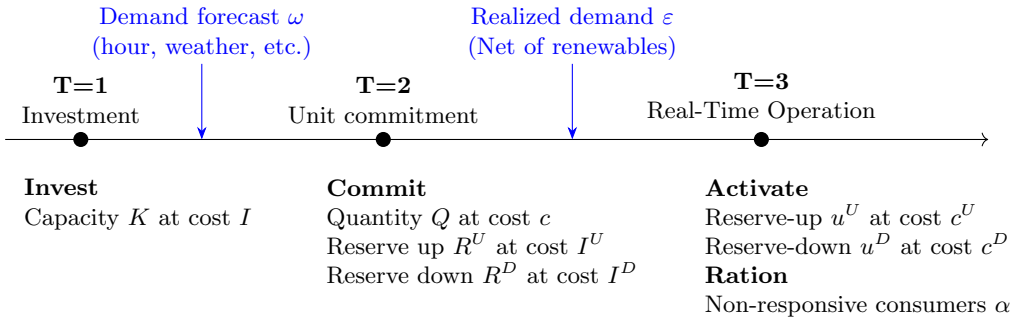


Figure 1: Timing of Decision and Information Revelation

3.2 Supply Side

To illustrate the flexibility characteristics of generating technologies in our model, we first introduce a generator with a single technology. We then extend to multiple technologies. We can describe the generators' actions in the three stages as indicated also in figure 1:

Stage 1 (Investment). The generator chooses the capacity level K, which determines the maximal production capacity in later stages.

Stage 2 (Unit Commitment). Based on forecast information ω, the generator commits to a production level Q_ω, upward reserves R^U_ω, and downward reserves R^U_ω, incurring corresponding commitment costs. The sum of scheduled production and upward reserves needs to be less than installed capacity, and the amount of

downward reserves must be less than committed production capacity:

$$R_\omega^D \leq Q_\omega \leq K - R_\omega^U. \tag{2}$$

Stage 3 (Real-Time Adjustment). Once uncertainty is resolved, a fraction u_ξ^U of the upward reserves and a fraction u_ξ^U of the downward reserves are activated, resulting in additional operating costs. The real-time production q_ξ is the scheduled production level Q_ω plus the fraction u_ξ^D of upward reserves R_ω^U minus the fraction u_ξ^D of downward reserves R_ω^D:

$$q_\xi = Q_\omega + u_\xi^U R_\omega^U - u_\xi^D R_\omega^D \qquad \text{with} \qquad u_\xi^U, u_\xi^D \in [0,1]. \tag{3}$$

The total cost incurred in uncertainty trajectory $\xi = (\omega, \varepsilon)$ is:

$$C_\xi = IK + cQ_\omega + (I^U + u_\xi^U c^U)R_\omega^U + (I^D - u_\xi^D c^D)R_\omega^D. \tag{4}$$

This equation summarizes the costs over all three stages. The first term represents the investment cost IK for installing capacity K in stage 1. The second term cQ_ω is the cost for scheduling production Q_ω in stage 2. The third and fourth terms represent the costs associated with scheduling and activating upward and downward reserves. The commitment costs in stage 2 for upward and downward reserves are $I^U R_\omega^U$ and $I^D R_\omega^D$ respectively. In stage 3, for activating upward reserves, the generator pays a per unit cost of c^U, while it recovers c^D per unit of output reduction. Hence, the minus sign for the downward reserves.

Together, equations 2-4 describe the generator's decision problem in managing flexibility: the sequence of actions and the association cost structure. We formally define a generation technology $\theta \in \Theta$ as a set of six cost parameters that represent investment, flexibility, and production characteristics. These parameters are summarized in Table 1. By varying cost parameters, the model can represent fully flexible, partially flexible, and fully inflexible technologies as special cases.

To avoid corner solutions and non-trivial commitment decisions, we impose the following assumptions.

Assumption 1 (Cost Parameter Structure)**.** *Each technology $\theta \in \Theta$ is defined by six non-negative cost parameters I_θ, c_θ, I_θ^U, I_θ^D, c_θ^U, c_θ^D ≥ 0, satisfying the following conditions:*

Table 1: Cost Parameters Defining a Technology θ

Symbol	Description
c_θ	Marginal cost of scheduled production
c_θ^U	Marginal cost of upward adjustment
c_θ^D	Marginal saving from downward adjustment
I_θ	Investment cost per unit of capacity
I_θ^U	Commitment cost per unit of upward reserves
I_θ^D	Commitment cost per unit of downward reserves

1. ***Marginal cost ordering:***

$$c_\theta^D \leq c_\theta \leq c_\theta^U.$$

It is more efficient to schedule production early, and adjusting production in real time comes at a cost. Downward adjustments allow partial recovery of production costs, but not fully, while upward adjustments are more expensive than scheduled production.

2. ***Reserve trade-off condition:***

$$c_\theta - c_\theta^U \leq I_\theta^U - I_\theta^D \leq c_\theta - c_\theta^D. \tag{5}$$

This ensures that both upward and downward reserves can be optimal depending on the probability distribution of demand.

The last condition can be understood by a simple example. Consider a generator that provides flexibility to satisfy random demand: low demand Q_L (with probability ϕ_L) and high demand Q_H (with probability $1 - \phi_L$). It can provide flexibility in two ways: (i) By using an **upward reserve strategy** and scheduling low output Q_L, and upward reserves $R^U = \Delta Q = Q_H - Q_L$ incurring expected costs.

$$cQ_L + (I^U + c^U(1 - \phi_L))\Delta Q.$$

(ii) By using a **downward reserve strategy** and scheduling high output Q_H and downward reserves $R^D = \Delta Q$, incurring cost

$$cQ_H + (I^D - c^D\phi_L)\Delta Q$$

The condition in equation (5) ensures that the upward reserves are preferred when

low demand is likely ($\phi_L \approx 1$), and downward reserves are preferred when high demand is likely ($\phi_L \approx 0$). Hence, both forms of flexibility can be optimal, depending on the shape of the demand distribution.

Note The commitment cost of downward reserves is often assumed to be zero ($I^D = 0$) as in Chao and Wilson (2002). We prefer a more general formulation, as it could account for richer interpretations. For instance, a positive I^D may reflect staffing costs to prepare for a production shutdown in response to real-time information.

The impact of real-time production decisions in stage 3 can be represented by a real-time supply curve as in Figure 2. The curve represents the marginal cost of adjusting output q_ξ, ignoring sunk investment and commitment costs. In our formulation this is a stepped function for each technology: the width of each step reflects the generator's commitment in Stage 2 (i.e. the levels of ($Q_{\theta,\omega}$, $R^U_{\theta,\omega}$, and $R^D_{\theta,\omega}$), and the height of the steps reflect marginal costs of short-term adjustments (c^U_θ and c^D_θ). A flatter real-time supply curve is associated with a more flexible generator, while a steeper curve represents inflexible generation. Flexibility in stage 3 is determined by both:

- How close the short-term adjustments costs c^U_θ and c^D_θ are to the scheduled marginal cost c, and
- The amount of upward and downward reserves $R^U_{\theta,\omega}$ and $R^D_{\theta,\omega}$ committed to in stage 2. A high reserve commitment requires lower commitment costs I^D_θ and I^U_θ.

Figure 2 illustrates this for three representative technology types:

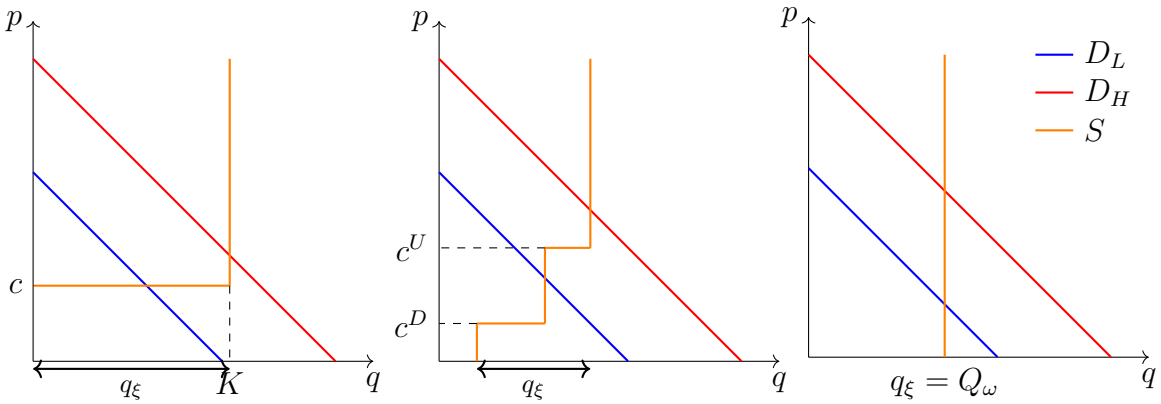
Panel (2a) represents the case where a technology is **perfectly flexible**, which corresponds to zero commitment costs $I^U = I^D = 0$ and real-time adjustments costs equal to the scheduled marginal production costs $c^U = c^D = c$. Since flexibility is free, the commitment stage becomes irrelevant, and the model reduces to a standard two-stage peak-load pricing model with investments and real-time operation. Generation firm can choose any production quantity q_ξ up to installed capacity K, depending on the realized state.

Panel (2c) shows the case when production technology is **non-responsive**, in which case real-time adjustment is impossible. This arises, for instance, for very high commitment costs for reserves $I^U = I^D = \infty$. This implies that the real-time production q_ξ must be equal to the pre-committed level $q_\xi = Q_\omega$. Absent flexibility, only the information ω available at stage 2 can be used to plan output.

Panel (2b) shows the case for a **limited flexible technology**. This case represents the general setting in our model. The firm commits to a quantity Q_ω, upward and downward reserve R^U_ω and R^D_ω. Real-time production $q_{\theta,\xi}$ in stage 3 must lie in the interval:

$$q_{\theta,\xi} \in [Q_\omega - R^D_\omega, Q_\omega + R^U_\omega].$$

Real-time flexibility exists but is costly: adjustments require prior commitment and incur non-zero marginal costs.



(a) Perfectly Flexible **(b)** Limited Flexible **(c)** Non-Responsive

Figure 2: Real-time Supply Functions of Three Technologies

We now extend the model to the set Θ of generation technologies. Each technology $\theta \in \Theta$ is characterized by its own cost parameters and decision variables. The total cost in state ξ is given by:

$$C_{\Theta,\xi} = \sum_{\theta \in \Theta} C_{\theta,\xi},$$

where $C_{\theta,\xi}$ is defined as in Equation 4 for each technology. We also define the aggregate real-time generation decisions across all technologies:

$$q_{\Theta,\xi} = \sum_{\theta \in \Theta} q_{\theta,\xi}.$$

3.3 Demand Side

We model consumer flexibility and rationing following the framework of Joskow and Tirole (2007). Before distinguishing between consumer types and their information sets, we first introduce the general setup and notation. Let $D(p)$ denote

the demand function, which gives the quantity consumers wish to consume at a price p, and let $S(p)$ be the associated *gross* consumer surplus. The surplus function is decreasing and concave in price.[12]

In some states, demand may exceed available supply and must be rationed. Let α denote the fraction of demand served (i.e., not rationed). The level of demand served is represented by the calligraphic letter $\mathcal{D}$ and given by:[13]

$$\mathcal{D}(p, \alpha) = \alpha D(p).$$

Let the function $\mathcal{S}(p, \alpha)$ denote the gross consumer surplus when the price is p and a fraction α of demand is served. The shape of this function depends on the rationing technology. Under random rationing, we have:

$$\mathcal{S}^{\text{Rand}}(p, \alpha) = \alpha S(p). \tag{6}$$

Smart meters may make rationing more efficient by rationing lower-valued consumption first. Under under efficient rationing, the expression becomes:

$$\mathcal{S}^{\text{Eff}}(p, \alpha) = S(p^*) \qquad \text{with} \qquad D(p^*) = \alpha D(p). \tag{7}$$

Following Borenstein and Holland (2007), consumers are divided into two categories: a fraction σ is price-insensitive, and the remaining fraction $(1-\sigma)$ is price-sensitive. The parameter σ is taken as exogenous. Price-sensitive consumers adjust their consumption in response to real-time information ξ, but price-insensitive consumers can only adjust consumption based on the price information available in the forward state ω.

The served demand and gross surplus are defined as follows. For price-insensitive consumers:

$$\mathcal{D}_\xi(p_\omega, \alpha_\xi), \quad \mathcal{S}_\xi(p_\omega, \alpha_\xi), \tag{8}$$

and for price-sensitive consumers :

$$\hat{\mathcal{D}}_\xi(p_\xi, \hat{\alpha}_\xi), \quad \hat{\mathcal{S}}_\xi(p_\xi, \hat{\alpha}_\xi), \tag{9}$$

where α_ξ and $\hat{\alpha}_\xi$ denote the fraction of demand served, p_ω is the price paid by

[12] The function $S(p)$ is the area under the demand curve $D(p)$. The two functions are related through the identity $S'(p) = pD'(p)$ and are both zero $S(\bar{p}) = D(\bar{p}) = 0$ at the reservation price $\bar{p}$.

[13] Joskow and Tirole (2007) argue that this relation does not hold, $\mathcal{D} \neq \alpha D$, if consumers adjust their behavior in response to higher levels of rationing, for instance by avoiding elevators.

price-insensitive consumers (set in advance), and p_ξ is the real-time price faced by price-sensitive consumers.

We defined the *value of lost load* (VOLL) as the marginal social value of restoring curtailed demand:

$$VOLL_\xi = \frac{\partial \mathcal{S}_\xi}{\partial \alpha_\xi} / \frac{\partial \mathcal{D}_\xi}{\partial \alpha_\xi}. \tag{10}$$

3.4 Social Optimum

The social optimum requires choosing investment capacity K_θ, commitments for energy $Q_{\theta,\omega}$, upward reserves $R^U_{\theta,\omega}$, downward reserves $R^D_{\theta,\omega}$, the reserves' utilization rates $u^U_{\theta,\xi}$, $u^D_{\theta,\xi}$, the price p_ω, p_ξ for price-insensitive and price-sensitive consumers respectively, and their rationing rates α_ξ, $\hat{\alpha}_\xi$, to maximize the social welfare function:

$$\begin{aligned}
\max \quad & \mathbb{E}_\xi\Big\{\sigma \mathcal{S}_\xi(p_\omega, \alpha_\xi) + (1-\sigma)\hat{\mathcal{S}}_\xi(p_\xi, \hat{\alpha}_\xi) - C_{\Theta,\xi}\Big\}, && (11)\\
\text{s.t.} \quad & \sigma \mathcal{D}_\xi(p_\omega, \alpha_\xi) + (1-\sigma)\hat{\mathcal{D}}_\xi(p_\xi, \hat{\alpha}_\xi) \le q_{\Theta,\xi} && [\eta_\xi h_\xi]\\
& Q_{\theta,\omega} + R^U_{\theta,\omega} \le K_\theta && [\lambda_{\theta,\omega}]\\
& R^D_{\theta,\omega} \le Q_{\theta,\omega} && [\mu_{\theta,\omega}]\\
& Q_{\theta,\omega} \ge 0 && [\varphi^Q_{\theta,\omega}]\\
& R^D_{\theta,\omega} \ge 0 && [\varphi^D_{\theta,\omega}]\\
& R^U_{\theta,\omega} \ge 0 && [\varphi^U_{\theta,\omega}]\\
& u^U_{\theta,\xi}, u^D_{\theta,\xi} \in [0,1].
\end{aligned}$$

Let $\eta_\xi h_\xi$ denote the multiplier of energy balance in real-time state ξ, $\lambda_{\theta,\omega}$ the capacity value of technology θ in state ω, and $\mu_{\theta,\omega}$ the energy value of technology θ in state ω. The last three inequalities are non-negative constraints.

Proposition 1. *Given cost parameters and state distributions $g(\omega), \phi(\varepsilon|\omega)$, the first-order conditions representing the optimum of maximization problem (11) are*

(a) Price-sensitive consumers:

No rationing

$$\forall \xi \qquad \hat{\alpha}_\xi = 1. \tag{12}$$

Price equal to energy shadow price

$$\forall \xi \qquad p_\xi = \eta_\xi. \tag{13}$$

(b) Price-insensitive consumers:

Optimal rationing

$$\forall \xi \qquad VOLL_\xi = \frac{\partial \mathcal{S}_\xi}{\partial \alpha_\xi} / \frac{\partial \mathcal{D}_\xi}{\alpha_\xi} = p_\xi \quad or \quad \alpha_\xi \in \{0, 1\}. \tag{14}$$

Retail price p_ω balances deadweight loss

$$\forall \omega \qquad \mathbb{E}_{\xi|\omega} \left[\frac{\partial \mathcal{S}_\xi}{\partial p_\omega} - p_\xi \frac{\partial \mathcal{D}_\xi}{\partial p_\omega} \right] = 0. \tag{15}$$

(c) Efficient activation of reserves.

$$\forall \xi, \theta \quad u^D_{\theta,\xi} = \begin{cases} 1, & p_\xi < c^D_\theta \\ \in [0,1], & p_\xi = c^D_\theta \\ 0, & p_\xi > c^D_\theta \end{cases} \qquad u^U_{\theta,\xi} = \begin{cases} 1, & p_\xi > c^U_\theta \\ \in [0,1], & p_\xi = c^U_\theta \\ 0. & p_\xi < c^U_\theta \end{cases} \tag{16}$$

(d) Commitment.

Energy commitments are like forward contracts

$$\pi^Q_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta \qquad = \lambda_{\theta,\omega} - \mu_{\theta,\omega} - \varphi^Q_{\theta,\omega}. \tag{17}$$

Downward reserves are like put options

$$\pi^D_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[\max\{c^D_\theta - p_\xi, 0\}] - I^D_\theta \qquad = \mu_{\theta,\omega} - \varphi^D_{\theta,\omega}. \tag{18}$$

Upward reserves are like call options

$$\pi^U_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[\max\{p_\xi - c^U_\theta, 0\}] - I^U_\theta \qquad = \lambda_{\theta,\omega} - \varphi^U_{\theta,\omega}. \tag{19}$$

(e) Capacity Investment.

$$\begin{aligned} &\mathbb{E}_{\omega}[\lambda_{\theta,\omega}] = I_{\theta} \quad \textit{if} \quad K_{\theta} > 0 \\ &\textit{otherwise, } K_{\theta} = 0. \end{aligned} \tag{20}$$

We interpret the first-order conditions as follows: Proposition 1 (a) indicates that price-sensitive consumers should face the real-time price p_{ξ} and are never rationed, as in Joskow and Tirole (2007). The intuition is straightforward that price-sensitive consumers can fully adjust their consumption according to state-contingent prices, and any curtailment is a dead-weight loss.

Proposition 1 (b) gives the optimal conditions for price-insensitive consumers. First, in case of rationing, the optimal rationing rate in state ξ should be such that VOLL in that state VOLL_{ξ} is equal to the real-time price p_{ξ}. Note that the real-time price is equal to the shadow price of the production constraint and reflects the social value of energy in state ξ.

The retail price p_{ω} that consumers pay in state ω is such that the marginal deadweight losses across states ξ are minimized. When there is no rationing, $\alpha_{\xi} = 1$

With random rationing, the optimal retail price that consumers pay is equal to the expected real-time price plus a correction reflecting the covariance between the real-time price and the normalized demand slope across states ξ:

$$p_{\omega} = \mathbb{E}_{\xi|\omega}(p_{\xi}) + \frac{\mathbb{C}ov_{\xi|\omega}\left[p_{\xi}, \alpha_{\xi}\partial D_{\xi}/\partial p_{\omega}\right]}{\mathbb{E}_{\xi|\omega}(\alpha_{\xi}\partial D_{\xi}/\partial p_{\omega})}.$$

Proposition 1 (c) shows that for scheduled downward reserves, only technologies for which the marginal production cost is larger than the real-time price are activated. And for scheduled upward reserves, only technologies for which the marginal production cost is smaller than the real-time price are activated. Note that in real time, neither capacity investment nor flexibility commitment is reversible, so any associated costs are sunk. The marginal cost only depends on marginal production cost c_{θ}^{U} and c_{θ}^{D}, respectively.

Proposition 1 (d) gives the necessary first-order conditions in the scheduling stage ($T = 2$). The three expressions on the left-hand side of the equation indicate the values of committing to sell energy (π_{ω}^{Q}), downward reserves (π_{ω}^{D}), and upward reserves (π_{ω}^{U}). Committing to sell energy is like a forward contract, downward reserves are like a put option on the real-time price, and upward reserves are like a call option on the real-time price. Let $\pi^{c}(X, P_{\omega})$ and $\pi^{p}(X, P_{\omega})$ denote the

expected return on a call option and a put option with strike price X and option price P_ω. That is

$$\pi^c_\omega(X, P_\omega) = \mathbb{E}_{\xi|\omega}[\max\{p_\xi - X, 0\}] - P_\omega. \tag{21}$$

$$\pi^p_\omega(X, P_\omega) = \mathbb{E}_{\xi|\omega}[\max\{X - p_\xi, 0\}] - P_\omega. \tag{22}$$

then $\pi^D_\omega = \pi^{put}(c^D, I^D)$ and $\pi^U_\omega = \pi^{call}(c^U, I^U)$. However, the commitment decisions are not financial contracts that can be sold independently; instead, they need to be backed up by generation capacity. There are the following inter-linkages: (1) In order to provide downward reserves, a generator needs to be scheduled to run in the first place. (2) A generation capacity can provide energy or upward reserves, but not both at the same time. Those linkages are described through the Lagrange multipliers on the right side of the equation.

The social planner will use the capacity of technology θ in state ω to maximize its social value $\lambda_{\theta,\omega}$, which is given by:

$$\lambda_{\theta,\omega} = \max\left\{0,\ \pi^U_{\theta,\omega},\ \pi^Q_{\theta,\omega} + \max\left\{\pi^D_{\theta,\omega}, 0\right\}\right\}. \tag{23}$$

The social planner can decide (1) to keep capacity idle and receive a value 0, (2) commit to upward reserves and make $\pi^U_{\theta,\omega}$, (3a) schedule energy and receive $\pi^Q_{\theta,\omega}$, or (3b) schedule energy and downward reserves $\pi^Q_{\theta,\omega} + \pi^D_{\theta,\omega}$.

Proposition 1(e) is the standard free-entry condition for investment of technology θ. Each investment opportunity earns zero profit. If the expected profit is negative because of a cost disadvantage, there is no investment in this technology.

3.5 Decentralized Market

Definition 1. *Let x_θ be the vector of decision variables of a firm with technology θ: production levels $q_{\theta,\xi}$, unit commitments $Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega}$, investments K_θ,*

$$x_\theta = (q_{\theta,\xi}, Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega}, K_\theta).$$

Let y be the vector of decision variables of a retailer selling to non-responsive consumers: a fixed fee $\mathcal{A}_\omega$, a retail price for non-responsive consumers p_ω, and the rationing rules α_ξ:

$$y = (\mathcal{A}_\omega, p_\omega, \alpha_\xi).$$

Let z be the vector of price variables: the real-time price p_ξ and the reservation utility of non-responsive consumers U_ω:

$$z = (p_\xi, U_\omega).$$

The vectors x^*_θ, y^*, z^* form a competitive equilibrium if

- *for a given z^*, x^*_θ maximizes the profit of a generator with technology θ,*
- *for a given z^*, y^* maximizes the retailers profit,*
- *in each state ξ all markets clear:*

$$q^*_{\Theta,\xi} = \sigma \mathcal{D}_\xi(p^*_\omega, \alpha^*_\xi) + (1-\sigma)\hat{\mathcal{D}}_\xi(p^*_\xi).$$

- *retailers make zero profit,*
- *the reservation utility of non-responsive consumers U_ω is consistent with the equilibrium contracts.*

Proposition 2. *The competitive equilibrium is equivalent to the social optimum.*

Proof. The profit maximization problem of a generator with technology θ is given by:

$$\begin{aligned} x^*_\theta \in \arg\max_{x_\theta} \mathbb{E}_\xi \left[p_\xi q_{\theta,\xi} - I_\theta K_\theta - c_\theta Q_\omega - (I^U_\theta + u^U_{\theta,\xi} c^U_\theta) R^U_{\theta,\omega} - (I^D_\theta - u^D_{\theta,\xi} c^D_\theta) R^D_{\theta,\omega} \right]. \\ \text{s.t.} \quad K_\theta \geq 0 \\ \forall \omega, \quad R^D_{\theta,\omega} \leq Q_{\theta,\omega} \leq K_\theta - R^U_{\theta,\omega} \quad \text{and} \quad Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega} \geq 0 \\ \forall \xi, \quad q_{\theta,\xi} = Q_{\theta,\omega} + u^U_\xi R^U_\omega - u^D_\xi R^D_\omega \quad \text{and} \quad u^U_\xi, u^D_\xi \in [0,1]. \end{aligned} \tag{24}$$

It is easy to verify that the first order conditions are given by the equation system from (c) to (e) in Proposition 1.

The retailer's profit maximization is given by:

$$\begin{aligned} y^* \in \arg\max_y \mathbb{E}_\xi [\mathcal{A}_\omega + (p_\omega - p_\xi)\mathcal{D}_\xi(p_\omega, \alpha_\xi)]. \\ \text{s.t. } \forall \omega, \quad \mathbb{E}_{\xi|\omega}\left[\mathcal{S}_\xi(p_\omega, \alpha_\xi) - p_\omega \mathcal{D}_\xi(p_\omega, \alpha_\xi) - \mathcal{A}_\omega\right] \geq U_\omega. \end{aligned} \tag{25}$$

Let ζ_ω be the multiplier of the constraint. The first order conditions of the retailers

are:

$$\zeta_\omega = 1$$
$$\mathbb{E}_{\xi|\omega}\left\{\mathcal{D}_\xi + (p_\omega - p_\xi)\frac{\partial D_\xi}{\partial p_\omega} + \zeta_\omega\left[\frac{\partial S}{\partial p_\omega} - D_\xi - p_\omega\frac{\partial D_\xi}{\partial p_\omega}\right]\right\} = 0$$
$$(p_\omega - p_\xi)\frac{\partial D_\xi}{\partial \alpha_\xi} + \zeta_\omega\left[\frac{\partial S}{\partial \alpha_\xi} - p_\omega\frac{\partial D_\xi}{\partial \alpha_\xi}\right] = 0.$$

Those equations are equivalent to the first-order conditions of the social planner characterized by (b). The zero profit condition determines the fixed part $\mathcal{A}_\omega$:

$$\mathcal{A}_\omega = \mathbb{E}_{\xi|\omega}[(p_\xi - p_\omega)\mathcal{D}_\xi(p_\omega, \alpha_\xi)].$$

This implies that consumers pay a fixed fee $\mathcal{A}_\omega$ if the volume-weighted real-time price p_ξ differs from the contract price p_ω. □

3.6 Value of Flexibility

The trade-off between commitment alternatives depends on the commitment costs, adjustment costs, and state distribution. Lemma 1 derives expressions for the value of generation capacity for two extreme cases: perfectly flexible generation and non-responsive generation, and shows that the value of a limited flexible technology is between those two extremes.

Lemma 1. *For the perfectly flexible and non-responsive technologies, the value of technology θ is given by:*

$$\lambda^{flex}_{\theta,\omega} = \mathbb{E}_{\xi|\omega}\{\max\{p_\xi - c_\theta, 0\}\} \tag{26}$$
$$\lambda^{nr}_{\theta,\omega} = \max\{\mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta, 0\}. \tag{27}$$

The value of a limited flexible technology is smaller than that of a completely flexible technology and larger than that of a non-responsive technology with the same marginal costs c_θ:

$$\lambda^{flex}_{\theta,\omega} \geq \lambda_{\theta,\omega} \geq \lambda^{nr}_{\theta,\omega}.$$

Proof. We need to calculate $\lambda_{\omega,\theta}$ in Equation 23. (i) For the perfectly flexible technology, there is no commitment cost for downward reserves, $I_D = 0$, so it is always optimal to commit to downward reserves as $\pi^D_{\theta,\omega} \geq 0$. The social planner has the choice between committing upward reserves, fully scheduling production, and committing to downward reserves, or leaving capacity idle. Given that $c^D =$

$c^U = c$ and there are no reservation costs, the two reserve strategies give the same profits $\pi^U_{\theta,\omega} = \pi^Q_{\theta,\omega} + \pi^D_{\theta,\omega}$. Those profits are weakly positive, so it is always optimal to provide flexibility. This proves equation 26. (ii). For the non-responsive technology, reserves are prohibitively expensive $\pi^U_{\theta,\omega} < 0$ and $\pi^D_{\theta,\omega} < 0$, so $\lambda^{\text{nr}}_{\theta,\omega} = \pi^Q_{\theta,\omega}$, which proves equation 27. □

Comparing the value of a limited flexible technology with that of a non-responsive technology, we can define a flexibility premium $\nu_{\theta,\omega}$.

Definition 2. *The flexibility premium for technology θ in state ω is*

$$\nu_{\theta,\omega} \equiv \lambda_{\theta,\omega} - \lambda^{nr}_{\theta,\omega}.$$

The flexibility premium ν measures the additional value a flexible asset will create compared to a non-responsive asset with the same marginal cost c_θ. By design, the flexibility premium is non-negative. The flexibility premium can compensate for flexible power plants with higher investment costs.

The next lemma shows that the flexibility premium ν in the forward market (commitment stage) can be seen as a combination of put and call options on the real-time price p_ξ.

Lemma 2. *The flexibility premium is the maximum of two put options when the expected real-time price is larger than c and the maximum of two call options otherwise, but can never become negative:*

$$\nu_{\theta,\omega} = \begin{cases} \max\{0,\ \pi^p_\omega(c^D, I^D),\ \pi^p_\omega(c^U, I^U + c^U - c)\} & \text{if } \mathbb{E}_{\xi|\omega}[p_\xi] > c \\ \max\{0,\ \pi^c_\omega(c^U, I^U),\ \pi^c_\omega(c^D, I^D + c - c^D)\} & \text{if } \mathbb{E}_{\xi|\omega}[p_\xi] \leq c \end{cases} \tag{28}$$

Proof. Using the definition of the flexibility premium, we can consider two cases. In case the non-responsive technology has a positive value $\lambda^{\text{nr}}_{\theta,\omega} = \pi^Q_{\theta,\omega} > 0$, the flexibility premium reflects the fact that (1) committing to downward reserves on top of the forward contract might provide the extra value $\pi^D_{\theta,\omega}$, (2) replacing the forward commitment with upward reserves might be more profitable: $\pi^U_{\theta,\omega} > \pi^Q_{\theta,\omega}$. So in that case we have:

$$\nu_{\theta,\omega} = \max\{0\ ,\ \pi^D_{\theta,\omega}\ ,\ \pi^U_{\theta,\omega} - \pi^Q_{\theta,\omega}\} \ \text{ if } \ \pi^Q_{\theta,\omega} > 0.$$

In case selling technology has no value, $\lambda^{\text{nr}}_{\theta,\omega} = 0$ and $\pi^Q_{\theta,\omega} \leq 0$, it could still be profitable to (1) schedule upward reserves instead, or (2) to schedule energy

forward only in combination with downward reserves.

$$\nu_{\theta,\omega} = \max\{0\ ,\ \pi^U_{\theta,\omega}\ ,\ \pi^Q_{\theta,\omega} + \pi^D_{\theta,\omega}\} \ \text{ if } \ \pi^Q_{\theta,\omega} < 0.$$

Above we have shown that $\pi^U_{\theta,\omega}$ and $\pi^D_{\theta,\omega}$ can be interpreted as call and put options. Using the call-put parity

$$\pi^c_\omega(X, A) - \pi^p_\omega(X, B) = \mathbb{E}_{\xi|\omega}[p_\xi] - X + B - A.$$

The two remaining terms $\pi^Q_{\theta,\omega} + \pi^D_{\theta,\omega}$ and $\pi^U_{\theta,\omega} - \pi^Q_{\theta,\omega}$ can also be written as options. For instance:

$$\pi^Q_{\theta,\omega} + \pi^D_{\theta,\omega} = \mathbb{E}_{\xi|\omega}(p_\xi) - c + \pi^p_\omega(c^D, I^D) = \pi^c_\omega(c^D, I^D + c - c^D).$$

□

Let $F_\omega(p_\xi)$ be the cumulative density function of the price distribution given information ω.

Lemma 3. *If F_{ω_1} second-order stochastically dominates F_{ω_2}, $F_{\omega_1} \succeq_{SSD} F_{\omega_2}$, and have the same expected value, then the flexibility premium in ω_1 is not larger than in state ω_1:*

$$\nu_{\theta,\omega_1} \leq \nu_{\theta,\omega_2}.$$

Proof. The value of call and put options increases when the tail of the distribution becomes more important or

$$\pi^{\mathrm{C}}_{\omega_1}(X, P) = \int_{\underline{p}}^{X} (F_{\omega_1}(p) - 1)dp - P \leq \int_{\underline{p}}^{X} (F_{\omega_2}(p) - 1)dp - P = \pi^{\mathrm{C}}_{\omega_2}(X, P), \quad (29)$$

$$\pi^{\mathrm{P}}_{\omega_1}(X, P) = \int_{\underline{p}}^{X} F_{\omega_2}(p)dp - P \leq \int_{\underline{p}}^{X} F_{\omega_1}(p)dp - P = \pi^{\mathrm{P}}_{\omega_2}(X, P). \quad (30)$$

The inequality follows from the definition: distribution $F_{\omega_1}(x)$ has second-order stochastic dominance over $F_{\omega_2}(x)$ if and only if $\int_{\underline{p}}^{p^*} F_{\omega_1}(x)dx \leq \int_{\underline{p}}^{p^*} F_{\omega_2}(x)dx$. This means that each term in the two parts of equation 28 is lower with ω_1. □

3.6.1 Example

We illustrate the decision procedure with a simple numerical example. A firm with a single unit of technology θ needs to commit to energy and reserves in state $\omega \in \{1, 2, 3, 4\}$, given that real-time prices can be either $p_L = 0$ or $p_H = 4$ with

probabilities $\phi(L|\omega)$ and $\phi(H|\omega)$. The technological parameters of technology θ are summarized in Table 2. The optimal commitment strategy is found by calculating π^Q, π^D, and π^U and selecting the option that maximizes the value of capacity using Equation 23. The results are shown in Table 3. In this example, commitment decisions change as the distribution over high and low prices changes. Energy commitment occurs if prices are likely to be high. With more uncertainty, energy commitment and downward reserves are optimal. If high prices are less likely, upward reserves are more attractive. If high prices are very unlikely, it might be better not to commit to the unit at all. If each state ω occurs with probability 1/4 the total value is $E_\omega(\lambda_\omega) = (2 + 21/20 + 3/16)/4 = 259/320$.

Parameter	**Value**
Marginal cost	$c = 1$
Upward adjustment cost	$c^U = \frac{5}{4}$
Downward adjustment cost	$c^D = \frac{1}{2}$
Upward commitment cost	$I^U = \frac{1}{2}$
Downward commitment cost	$I^D = \frac{1}{5}$
Firm's capacity	$K = 1$

Table 2: Technological Assumptions

ω	$\phi(L\|\omega)$	$\phi(H\|\omega)$	p_F	π^Q	π^D	π^U	λ_ω	Commitment	ν
1	$\frac{1}{4}$	$\frac{3}{4}$	3	2	$-\frac{3}{40}$	$\frac{25}{16}$	2	Q	0
2	$\frac{1}{2}$	$\frac{1}{2}$	2	1	$\frac{1}{20}$	$\frac{7}{8}$	$\frac{21}{20}$	Q and R^D	$\frac{1}{20}$
3	$\frac{3}{4}$	$\frac{1}{4}$	1	0	$\frac{7}{40}$	$\frac{3}{16}$	$\frac{3}{16}$	R^U	$\frac{3}{16}$
4	$\frac{9}{10}$	$\frac{1}{10}$	$\frac{2}{5}$	$-\frac{3}{5}$	$\frac{1}{4}$	$-\frac{9}{40}$	0	None	0

Table 3: Profitability and Optimal Strategy under Different States ω.

4 Effects of Market Design on Flexibility Investment

In this section, we first investigate the ability of different market structures to capture the flexibility premium. Then, we show the effects of reserve market design on flexibility investment and social welfare.

4.1 Forward Market without Real-Time Market

We now consider an environment in which all wholesale electricity is traded in a forward market before uncertainty is resolved. In contrast to the decentralized benchmark in Section 3.5, there is no wholesale real-time market: generators and retailers contract a fixed quantity at a single forward price p_ω^F, determined prior to the realization of demand. Once the state of demand ξ is known, contracted positions and production levels are fixed, and there is no further wholesale adjustment or re-pricing.

To isolate the effect of replacing the wholesale real-time market with a wholesale forward market, we retain very efficient retailers that balance demand with their contracted supply in each state of the world. These retailers use rich two-part contracts, which allow retail prices to reflect the social value of electricity while protecting them from opportunistic consumer switching in the absence of a short-term wholesale market price.

The retailer balances an internal energy constraint in each state of the world. Given its fixed forward position q_ω^R, total retail demand must equal this contracted supply in every realized state ξ. To achieve this balance, the retailer relies on two instruments: real-time pricing for responsive consumers and rationing for non-responsive consumers. Through these mechanisms, scarcity is efficiently allocated across consumer groups and states of the world, even though there is no wholesale market in which production or prices can adjust ex post. Two-part tariffs for both consumer groups ensure that these instruments can be applied efficiently while maintaining the retailer's zero-profit condition.

We will set up the model in a structure similar to the decentralized equilibrium in Section 3.5, but with some important changes:

(i) *Wholesale market.* Wholesale trading occurs only at the forward stage ω, where quantities are contracted at a forward price p_ω^F. There is no wholesale market in the real-time states ξ, and hence no price that reflects the social value of electricity ex post.

(ii) *Retail market.* Retailers maintain balance internally in each realized state ξ, based on their fixed wholesale positions q_ω^R. They adjust retail demand through real-time pricing for responsive consumers and rationing for non-responsive consumers.

(iii) *Firms' decisions.* Generators choose capacity K_θ and forward commitments $Q_{\theta,\omega}$ based on expected conditions. Reserve and activation variables

$(R^U_{\theta,\omega}, R^D_{\theta,\omega}, u^U_{\theta,\xi}, u^D_{\theta,\xi})$ remain feasible but yield no value when the wholesale price is fixed; in equilibrium, they are set to zero.

(iv) *Market clearing and prices.* Wholesale balance holds ex ante at the forward stage,

$$\sum_{\theta} Q_{\theta,\omega} = q^R_\omega,$$

while retail balance holds ex post within each retailer through its internal mechanisms $(\hat{p}_\xi, \alpha_\xi)$. In equilibrium, the expected internal retail price will be equal to the forward price,

$$p^F_\omega = \mathbb{E}_{\xi|\omega}[\hat{p}_\xi].$$

Implications. Allowing retail-side adjustment mitigates inefficiencies on the demand side (responsive consumers still see a real-time price signal $\hat{p}_\xi$), but flexible *generation* cannot monetize variation in system value because the wholesale price is fixed at p^F_ω. The flexibility premium for supply therefore vanishes, leading to distorted investment in flexible technologies even under an efficient retailer.[14]

Definition 3 (Forward Market Equilibrium)**.** *Let x_θ be the vector of decision variables of a firm with technology θ: production levels $q_{\theta,\xi}$, unit commitments $Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega}$, investments K_θ,*

$$x_\theta = (q_{\theta,\xi}, Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega}, K_\theta).$$

Let y be the vector of decision variables of a retailer selling to non-responsive consumers: a fixed fee A_ω, a retail price for non-responsive consumers p_ω, and the rationing rule $\hat{\alpha}_\xi$, and decision variables of a retailer selling to responsive consumers: a fixed fee $\hat{A}_\xi$, a retail prices for responsive consumers $\hat{p}_\xi$, and the rationing rule α_ξ, and the quantity purchased in the forward market q^R_ω:

$$y = (\mathcal{A}_\omega, p_\omega, \alpha_\xi, \hat{\mathcal{A}}_\xi, \hat{p}_\xi, \hat{\alpha}_\xi, q^R_\omega)$$

Let z be the vector of price variables: the forward price p^F_ω, the reservation utility of non-responsive consumers U_ω and the reservation utility of responsive

[14] A less efficient and arguably more realistic alternative is to assume that, in the absence of a wholesale real-time market, the system operator applies non-price rationing (e.g., random curtailment) and retailers neither compete in rationing rules nor deploy state-dependent retail prices. Our results below then provide an *upper bound* on performance; the less efficient variant would strengthen the investment distortions.

consumers $\hat{U}_\xi$*:*

$$z = (p^F_\omega, U_\omega, \hat{U}_\xi)$$

The vectors x^*_θ, y^*, z^* *form a competitive equilibrium if*

- *for a given* z^**,* x^*_θ *maximizes the profit of a generator with technology* θ*,*
- *for a given* z^**,* y^* *maximizes the retailers profit,*
- *in each state* ξ *all markets clear:*

$$q^*_{\Theta,\xi} = q^*_{\Theta,\omega} = \sigma\mathcal{D}_\xi(p^*_\omega, \alpha^*_\xi) + (1-\sigma)\hat{\mathcal{D}}_\xi(\hat{p}^*_\xi, \hat{\alpha}^*_\xi).$$

- *retailers and generators make zero profit,*
- *the reservation utility of non-responsive consumers* U_ω *and of responsive consumers* $\hat{U}_\xi$ *is consistent with the equilibrium contracts.*

Now, we build the maximization problem of a generator with technology θ as:

$$\begin{aligned} x^*_\theta \in \arg\max_{x_\theta} \mathbb{E}_\xi \left[p^F_\omega q_{\theta,\xi} - I_\theta K_\theta - c_\theta Q_\omega - (I^U_\theta + u^U_{\theta,\xi}c^U_\theta)R^U_{\theta,\omega} - (I^D_\theta - u^D_{\theta,\xi}c^D_\theta)R^D_{\theta,\omega}\right] \\ \text{s.t.} \quad K_\theta \geq 0, \\ \forall\omega, \quad R^D_{\theta,\omega} \leq Q_{\theta,\omega} \leq K_\theta - R^U_{\theta,\omega} \quad \text{and} \quad Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega} \geq 0 \\ \forall\xi, \quad q_{\theta,\xi} = Q_{\theta,\omega} + u^U_{\theta,\xi}R^U_{\theta,\omega} - u^D_{\theta,\xi}R^D_{\theta,\omega} \quad \text{and} \quad u^U_{\theta,\xi}, u^D_{\theta,\xi} \in [0,1] \end{aligned} \tag{31}$$

and the retailer's profit maximization is given by:

$$\begin{aligned} y^* \in \arg\max_{y} \mathbb{E}_\xi\{\sigma[A_\omega + p_\omega\mathcal{D}_\xi(p_\omega, \alpha_\xi)] + (1-\sigma)[\hat{A}_\xi + \hat{p}_\xi\hat{\mathcal{D}}_\xi(\hat{p}_\xi, \hat{\alpha}_\xi)] - p^F_\omega q^R_\omega\} \\ \text{s.t.} \ \forall\omega, \quad \mathbb{E}_{\xi|\omega}\Big[\mathcal{S}_\xi(p_\omega, \alpha_\xi) - p_\omega\mathcal{D}_\xi(p_\omega, \alpha_\xi) - \mathcal{A}_\omega] \geq U_\omega \\ \forall\xi, \quad \mathbb{E}_\xi\Big[\hat{\mathcal{S}}_\xi(\hat{p}_\xi, \alpha_\xi) - \hat{p}_\xi\hat{\mathcal{D}}_\xi(\hat{p}_\xi, \hat{\alpha}_\xi) - \hat{\mathcal{A}}_\xi] \geq \hat{U}_\xi \\ \forall\xi, \quad \sigma\mathcal{D}_\xi(p_\omega, \alpha_\xi) + (1-\sigma)\hat{\mathcal{D}}_\xi(\hat{p}_\xi, \hat{\alpha}_\xi) = q^R_\omega \end{aligned} \tag{32}$$

Let κ_ξ be the Lagrange multiplier of the retailer's retail quantity constraint.

Proposition 3. *The first-order conditions representing the optimum of the only forward market are:*

(a) No rationing for responsive consumers and optimal rationing for non-responsive consumers is such that the VOLL is equal to the retail price charged to responsive consumers:

$$\hat{\alpha}_\xi = 1. \tag{33}$$

$$\forall \xi, \frac{\partial \mathcal{S}_\xi}{\partial \alpha_\xi} / \frac{\partial \mathcal{D}_\xi}{\alpha_\xi} = \hat{p}_\xi \quad or \quad \alpha_\xi \in \{0, 1\}. \tag{34}$$

(b) The retail price p_ω for non-responsive consumers is:

$$\forall \omega \qquad \mathbb{E}_{\xi|\omega} \left[\frac{\partial \mathcal{S}_\xi}{\partial p_\omega} - p_\omega^F \frac{\partial \mathcal{D}_\xi}{\partial p_\omega} \right] = 0. \tag{35}$$

(c) Forward prices equal the expected real-time retail price

$$p_\omega^F = \mathbb{E}_{\xi|\omega}[\hat{p}_\xi].$$

(d) The zero-profit condition determines the fixed parts $\mathcal{A}_\omega$ and $\hat{A}_\xi$:

$$A_\omega = (p_\omega^F - p_\omega)\, \mathbb{E}_{\xi|\omega}[\mathcal{D}_\xi(p_\omega, \alpha_\xi)],$$
$$\hat{A}_\xi = (p_\omega^F - \hat{p}_\xi)\, \hat{\mathcal{D}}_\xi(\hat{p}_\xi, \hat{\alpha}_\xi).$$

(e) No commitment to reserves:

$$R_{\theta,\omega}^U = 0, R_{\theta,\omega}^D = 0. \tag{36}$$

(f) Energy commitment:

$$p_\omega^F - c_\theta = \lambda_{\theta,\omega} - \varphi_{\theta,\omega}^Q. \tag{37}$$

(g) Capacity investment:

$$\begin{aligned} &\mathbb{E}_\omega[\lambda_{\theta,\omega}] = I_\theta \quad if \quad K_\theta > 0 \\ &Otherwise, K_\theta = 0 \end{aligned} \tag{38}$$

Discussion In equilibrium, reserves are not used because, under our assumptions, keeping upward reserves available and activating them in all states would be strictly more costly than producing the contracted quantity directly. The value of a technology in state ω is therefore determined solely by its ability to sell a fixed

production quantity in the forward market at the price p^F_ω. Formally, the unit return to capacity is given by

$$\lambda^F_{\theta,\omega} = \max\{p^F_\omega - c_\theta,\, 0\},$$

which is a similar payment that a completely inflexible technology would obtain in a real-time market. Hence, technologies with different degrees of operational flexibility receive identical expected revenues, and the flexibility premium that exists in the real-time market equilibrium disappears. Because forward prices are endogenously linked to investment choices, the resulting equilibrium supports a different long-run investment mix than in the decentralized real-time market.

Under random rationing, both high- and low-value non-responsive consumers are curtailed proportionally within each realized state ξ. Using $\mathcal{D}_\xi(p_\omega, \alpha_\xi) = \alpha_\xi\, \mathcal{D}(p_\omega)$ and the forward-market condition $\mathbb{E}_{\xi|\omega}[\hat{p}_\xi] = p^F_\omega$, the first-order condition for p_ω can be written as

$$p_\omega = p^F_\omega + \frac{\mathrm{Cov}_{\xi|\omega}(\hat{p}_\xi, \alpha_\xi)}{\mathbb{E}_{\xi|\omega}[\alpha_\xi]}.$$

Compared to the decentralized benchmark, the structure of this expression is similar, but the covariance term differs in its weighting. In the decentralized market, the corresponding term involves $\mathrm{Cov}_{\xi|\omega}\big(p_\xi, \alpha_\xi\, \partial D_\xi/\partial p_\omega\big)$, reflecting that price adjustments there are weighted by the marginal responsiveness of demand. Under proportional rationing, these slope terms cancel out, so the covariance in the forward-only market depends only on the co-movement between the internal retail price $\hat{p}_\xi$ and the rationing share α_ξ.

The average of the state-contingent transfers $\hat{A}_\xi$ does not, in general, equal zero. From $\hat{A}_\xi = (p^F_\omega - \hat{p}_\xi)\, \hat{\mathcal{D}}_\xi$ and $\mathbb{E}_{\xi|\omega}[\hat{p}_\xi] = p^F_\omega$, we obtain

$$\mathbb{E}_{\xi|\omega}[\hat{A}_\xi] = -\,\mathrm{Cov}_{\xi|\omega}(\hat{p}_\xi, \hat{\mathcal{D}}_\xi).$$

Because flexible demand declines in high-price states, this covariance is negative, so the expected fixed payment $\mathbb{E}_{\xi|\omega}[\hat{A}_\xi]$ is positive. This positive transfer is not a reward for providing flexibility, but a financial adjustment that offsets the mismatch between the fixed wholesale forward price and the varying retail revenues in different states of demand.

4.2 Reserve Market

We extend the forward only model by introducing reserve markets. At the forward stage ω, generators and retailers trade forward contracts for certain delivery and reserve contracts that take the form of call options on energy with strike price $X \in \mathcal{X}$. The set $\mathcal{X} = \{c_\theta^U, c_\theta^D : \theta \in \Theta\}$ corresponds to the marginal adjustment costs of all technologies. We consider only call options, since put options that represent downward flexibility can be replicated by combining forward and call positions. Forward and option prices, p_ω^F and $\{p_\omega^X\}_{X\in\mathcal{X}}$, are determined endogenously in competitive equilibrium.

There is no real time wholesale market; all adjustments occur through the activation of reserve options by retailers. At the realization stage ξ, retailers balance demand using reserve activation, rationing for non-responsive consumers, and pricing for responsive consumers. Generators supply their contracted forward quantities and any activated reserves.

Definition 4 (Reserve-Market Equilibrium)**.** *A* reserve-market equilibrium *is a collection*

$$(x_\theta^*,\ y^*,\ z^*,\ e^*),$$

where the price vector z collects all state-contingent variables across forward and real-time states:

$$z = (p_\omega^F,\ p_\omega^X,\ p_\omega,\ U_\omega^{NR},\ U_\xi^R),$$

with p_ω^F is the wholesale forward price, p_ω^X the option price with strike price $X \in \mathcal{X}$, p_ω the retail price for non-responsive consumers, U_ω^{NR} and U_ξ^R the reservation utilities of non-responsive and responsive consumers,
where x_θ represents firm θ's decisions:

$$x_\theta = (K_\theta,\ Q_{\theta,\omega}, R_{\theta,\omega}^U, R_{\theta,\omega}^D, S_{\theta,\omega}^F, S_{\theta,\omega}^X,\ u_{\theta,\xi}^U, u_{\theta,\xi}^D)$$

with K_θ the investment levels, $Q_{\theta,\omega}$, $R_{\theta,\omega}^U$, $R_{\theta,\omega}^D$ day-ahead unit-commitment levels, $u_{\theta,\xi}^U$, $u_{\theta,\xi}^D$ reserve activation levels and $S_{\theta,\omega}^F$ and $S_{\theta,\omega}^X$ the sales of forwards and call options with strike price X,
where the vector y represents the decisions of a representative retailer in the market

$$y = (B_\omega^F, B_\omega^X, p_\omega,\ e_\xi^X, \hat{p}_\xi, A_\omega^{NR}, A_\xi^R, \alpha_\xi^{NR}, \alpha_\xi^R)$$

with B_ω^F and B_ω^X the contracts procured in the day-ahead markets, p_ω the retail price for non-responsive consumers, $\hat{p}_\xi$ the real price for responsive consumers,

A_ω^{NR} and A_ξ^R the fixed charges for non-responsive and responsive consumers, and α_ξ^{NR} and α_ξ^R the rationing of consumers in state ξ.
and the vector e represents the retailer's activation decisions $e_\xi^X \in [0,1]$ of the call option with strike price X in state ξ.

(1) Generators maximize profit. *A firm with technology $\theta \in \Theta$ maximizes expected profit given prices z^* and option activation levels e^**

$$x_\theta^* = \arg\max_{x_\theta} \ \mathbb{E}_\omega \mathbb{E}_{\xi|\omega}\Big[\Pi_{\theta,\omega,\xi}(x_\theta; z^*, e^*)\Big],$$

by choosing x_θ subject to dispatch constraints, the physical production constraint and the obligation for production to fulfill contract oblibations:

$$R_{\theta,\omega}^D \leq Q_{\theta,\omega} \leq K_\theta - R_{\theta,\omega}^U, \tag{39}$$

$$q_{\theta,\xi} = Q_{\theta,\omega} + u_{\theta,\xi}^U R_{\theta,\omega}^U - u_{\theta,\xi}^D R_{\theta,\omega}^D, \qquad u_{\theta,\xi}^U, u_{\theta,\xi}^D \in [0,1], \tag{40}$$

$$q_{\theta,\xi} = S_{\theta,\omega}^F + \sum_{X\in\mathcal{X}} e_\xi^{*X} S_{\theta,\omega}^X. \tag{41}$$

(2) Retailers maximize profit. *Retailers maximize expected profit by setting y and e subject to the consumer's participation constraints of the two consumer groups and the physical and contractual balancing constraints:*

$$U_\omega^{NR} \ \leq \mathbb{E}_{\xi|\omega}\Big[S_\xi(p_\omega, \alpha_\xi^{NR}) - p_\omega D_\xi(p_\omega, \alpha_\xi^{NR}) - A_\omega^{NR}\Big], \tag{42}$$

$$U_\xi^R \leq \hat{S}_\xi(\hat{p}_\xi, \alpha_\xi^R) - \hat{p}_\xi \hat{D}_\xi(\hat{p}_\xi, \alpha_\xi^R) - A_\xi^R, \tag{43}$$

$$q_\xi^{retail} = \sigma\, D_\xi(p_\omega, \alpha_\xi^{NR}) + (1-\sigma)\, \hat{D}_\xi(\hat{p}_\xi, \alpha_\xi^R), \tag{44}$$

$$q_\xi^{retail} = B_\omega^F + \sum_{X\in\mathcal{X}} e_\xi^X B_\omega^X. \tag{45}$$

(3) Markets clear. *There is a market equilibrium in the day-ahead market for the forward and all option markets, and the total level of physical production equals physical consumption*

$$\sum_\theta S_{\theta,\omega}^F = B_\omega^F, \qquad \sum_\theta S_{\theta,\omega}^X = B_\omega^X \tag{46}$$

$$\sum_\theta q_{\theta,\xi} = q_\xi^{retail} \tag{47}$$

(4) Free entry and consistency *The representative retailers make zero profit and the reservation utilities of consumers correspond to the utility they can obtain in equilibrium.*

We can now derive the following proposition:

Proposition 4. *The reserve market equilibrium obtains the first best outcome. The retailer will exercise call options iff their strike price X is less than the real-time retail price $\hat{p}_\xi$,*

$$e_\xi^X = \mathbf{1}_{\{X<\hat{p}_\xi\}} \quad \textit{with } e_\xi^X \in [0,1] \textit{ if } X = \hat{p}_\xi.$$

The day ahead forward price and option price are given by

$$\begin{aligned} p_\omega^F &= \mathbb{E}_{\xi|\omega}(\hat{p}_\xi), \\ p^X &= \mathbb{E}_{\xi|\omega}\{\max\{\hat{p}_\xi - X, 0\}\}. \end{aligned}$$

which follows from the retailer arbitrage conditions of buying call options and balancing demand and supply with price instruments and rationing. Generators will self-select in selling contracts that correspond to their physical positions and the marginal adjustment costs of their technology:

$$\begin{aligned} S_{\theta,\omega}^F &= Q_{\theta,\omega}, \\ S_{\theta,\omega}^X &= R_\theta^U \, \mathbf{1}_{\{X=c_\theta^U\}} + R_\theta^D \, \mathbf{1}_{\{X=c_\theta^D\}}. \end{aligned}$$

The generator's dispatch maximizes profits by choosing between not producing, committing to a fixed quantity, selling upward reserves, or committing to a fixed quantity and downward reserves.

$$\lambda_{\theta,\omega} = \max\{0, p_\omega^F - c_\theta, p_\omega^{c_\theta^U} - I_\theta^U, p_\omega^{c_\theta^D} - I_\theta^D - c_\theta + c_\theta^D\}$$

Free entry will drive down the investment decisions of all technologies so:

$$\mathbb{E}[\lambda_{\theta,\omega}] = I_\theta$$

Chao and Wilson (1987) show that in a retail market, the social optimum can be implemented equivalently through spot price or an array of incentive-compatible contingent contracts from which consumers self-select their own contract by their willingness to pay, which is privately known. Also, Oren (2003) claims the equivalence of spot pricing and technology-specific energy and capacity payment to recoup investment costs. This paper, by contrast, demonstrates that a real-time market can be replaced by a forward market, combined with a reserves market with a menu of call options.

4.3 Example

This section gives an example to quickly understand the key points of the only forward market and reserve market. There is one state in second stage ($|\Omega| = 1$) with two possible states: low demand: $\varepsilon = L$ with probability f_L, and high demand: $\varepsilon = H$ with probability f_H, $f_H + f_L = 1$; two technologies are available: $\theta \in \{1, 2\}$. Both technologies have the same production cost, $c_1 = c_2 = c$. Technology 1 is totally flexible, so it can postpone production until the state of the world is realized; technology 2 is totally inflexible, so it must commit itself to production before demand is known and is not able to adjust in real-time. That is, $I_1^U = I_1^D = 0, c_1^U = c_1^D = c_1$, and $I_2^U = I_2^D = c_2^U = c_2^D = \infty$[15]. The investment cost of technology 1 is larger than that of technology 2: $I_1 > I_2$. All consumers are responsive to real-time prices ($\sigma = 0$). The total capacity of each technology is denoted by K_1 and K_2.

Denote real-time prices for both states as p_L and p_H, respectively. The social planner's objective[16] is to to maximize social welfare by choosing capacity K_1, K_2, quantity committed by technology 2 before demand is realized Q_2, and quantity produced by technology 1 in low state $q_{1,L}$ and in high state $q_{1,H}$:

$$\begin{aligned} \max_{\{K_1,K_2,q_{1,\varepsilon},Q_2\}} & \mathbb{E}[S_\varepsilon(q_{1,\varepsilon} + Q_2) - c \cdot (q_{1,\varepsilon} + Q_2)] - I_1 K_1 - I_2 K_2. \\ s.t. \quad & q_{1,\varepsilon} \leq K_1, \\ & q_{2,\varepsilon} = Q_2, \\ & Q_2 \leq K_2. \end{aligned} \tag{48}$$

The first-order conditions yield:

$$\begin{aligned} p_H &= \frac{I_1}{f_H} + c, \\ p_L &= c - \frac{I_1 - I_2}{f_L} < c, \\ Q_2 &= K_2; q_1^H = K_1; q_1^L = 0, \\ K_2 &= S_L^{-1\prime}(p_L), \\ K_1 &= S_H^{-1\prime}(p_H) - K_2, \end{aligned}$$

Hence, in the presence of demand uncertainty and an efficient real-time market:

a) inflexible firms earn the expected price $E(p) = c + I_2$;

[15] By abuse of notation, we equalize the infinite numbers.

[16] As shown above, the competitive equilibrium is equivalent to the social planner solution.

b) flexible firms earn p_H in high demand state and do not produce in low demand state;

c) low demand price is below marginal production cost, $p_L < c$.

Flexible firms earn an expected premium equal to $I_1 - I_2$, which recoups the extra capacity investment cost of flexible assets. Now, consider the absence of a real-time market. Both types of firms need to determine the quantity and prices before demand is realized. The maximization problem becomes

$$\begin{aligned} \max_{\{K_1,K_2,Q_1,Q_2\}} & E[S_\varepsilon(Q_1+Q_2) - c\cdot(Q_1+Q_2)] - I_1K_1 - I_2K_2. \\ s.t. \quad & Q_1 \le K_1 \\ & Q_2 \le K_2. \end{aligned} \tag{49}$$

which gives

$$\begin{aligned} & K_1 = 0; Q_2 = K_2 > S_L^{-1\prime}(p_L), \\ & \mathbb{E}[S'(K_2)] = c + I_2, \\ & p_F = E_\varepsilon(p), \\ & p_L^R = D_L^{-1}(K_2), \\ & p_H^R = D_H^{-1}(K_2). \end{aligned}$$

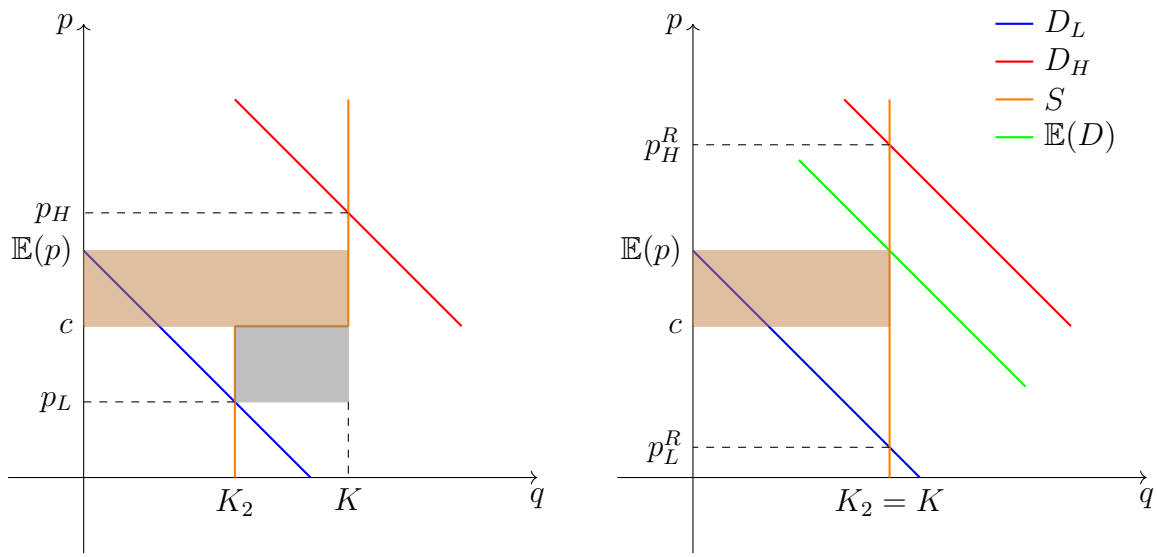
p_F is the unit payment to generators, and p_H^R, p_L^R are the retail price charged to consumers in high and low states, respectively. As predicted, in the absence of real-time markets and the presence of a forward market, long-term equilibrium shows:

a) no investment in flexible technology;

b) over-investment in inflexible technology;

In this case, there is a missing market problem; firms and consumers cannot directly trade on goods as $p_R < c$; they will not produce even though there is a fixed payment. Hence, there must be an intermediary (e.g., retailer) that pays p_F to the firm and charges a two-part tariff from consumers.

Next, we show how an options market complements the forward market and achieves an optimum even without a real-time market. Firms and consumers can trade in either a forward market or an options market, or both. Firms receive p_F



(a) Real-Time Market **(b)** Only Forward Market

Figure 3: Equilibrium in Real-time and Forward Market

Note: This figure represents the equilibrium comparison between a real-time market and a forward only market. Sub-figure (3a) shows the equilibrium prices, quantities, and investments when a well-functioning real-time market exists. In this case, firms making a production commitment earn the expected price $\mathbb{E}(p)$ and the net profit is shaded in brown. In addition, flexible firms can react to real-time demand shocks so they avoid the loss at the low state(shaded in gray). Flexibility premium is unconditional loss avoidance, so the brown area times the probability of being in a low state p_L, which is equal to $I_1 - I_2$. Sub-figure (3b) shows the equilibrium when a real-time market is not available. Commitment is required for both flexible and inflexible firms, according to expected demand, and flexible firms are not rewarded for providing flexibility. Hence, in the long term, no flexible assets survive. Furthermore, to balance the welfare loss from under-consumption in the low state and rationing in the high state. $K_2(b) > K_2(a)$.

for the unit quantity they sell in the forward market; if they trade in the options market, they receive a capacity payment p_K for being available in real-time and a production payment p_X if the option is activated. There is a huge penalty for not being available when activated to guarantee that only flexible firms enter the options market, and they will not sell more than they invest. The decision made by inflexible firms is trivial. They will produce up to capacity as long as the forward price p_F is higher than the production cost c. The objective function of flexible firms is

$$\max_{Q_1^F, Q_1^O, K_1} (p_F - c)Q_1^F + p_K Q_1^O + \mathbb{E}[(p_X - c)q_{1,\varepsilon}] - I_1 K_1.$$
$$s.t \quad q_{1,\varepsilon} \leq Q_1^O,$$
$$Q_1^F + Q_1^O \leq K_1.$$

Q_1^F is quantity sold in the forward market; Q_1^O is quantity sold in the options market; $q_{1,\varepsilon}$ is activated options in state ε. Social optimum can be attained through a forward market with forward price $p_F = E_\varepsilon(p)$ and an options market with capacity price p_K and strike price p_X described by:

$$p_K = f_H(p_H - p_X), \quad p_L \leq p_X \leq p_H. \tag{50}$$

As shown in Lemma 5, when there is only one technology providing flexibility, there are infinitely many combinations of strike and activation prices. If the strike price is equal to the production cost, $p_X = c$, a flexible firm should be paid a capacity payment p_K larger than the opportunity cost of not trading in the forward market:

$$p_K > \mathbb{E}_\varepsilon(p) - c. \tag{51}$$

Proof. When $p_X = c, p_K = f_H(p_H - c) = I_1 > I_2 = E_\varepsilon(p) - c.$ □

4.4 Existing Reserve Market Designs

4.4.1 Integrated Market for Energy and Reserves

Many markets in the U.S. advocate a co-procurement of energy and reserves, and compensate reserve providers for the opportunity cost of not selling in the energy market, which would lead to insufficient investment in flexible capacity, even without any direct cost associated with adjustment. Based on the idea of opportunity cost reimbursement, in an integrated auction, generators that produce receive the market-clearing price p_ω^F, and the plants providing unemployed reserves receive the difference between the forward price and their own bid $p_\omega^F - b_\theta$ as capacity payment, and the bid b_θ serves as the strike price. Given the cost information is perfectly known (Hogan 2005, 2013; Oren, 2003), the capacity payment is $p_\omega^F - c_\theta$, but if the cost is private information, the generators would bid below true cost and earn an information rent (Oren and Sioshansi, 2005). Hence, reserve providers whose capacity is deployed receive p_ω^F, while unemployed reserves are paid $p_\omega^F - b_\theta > p_\omega^F - c$. Assume the distribution of the forward clearing price is F_ω^F. A firm with production cost c_θ has expected profit[17]:

$$\pi_{\theta,\omega} = \int_{b_\theta}^{\infty} (p_\omega^F - b_\theta) dF_\omega^F + (b_\theta - c_\theta)(1 - F_\omega(b_\theta))(1 - F_\omega^F(b_\theta)) \tag{52}$$

[17]To keep matters simple and comparable to Oren and Sioshansi (2005), I drop direct costs of providing flexibility: $I_\theta^U = I_\theta^D = 0, c_\theta^U = c_\theta^D = c_\theta$. Adding up these costs does not alter the results.

The optimal bidding is given by:

$$b_\theta^o = c_\theta - \frac{[1 - F_\omega^F(b_\theta^o)]F_\omega(b_\theta^o)}{F_\omega(b_\theta^o)[1 - F_\omega^F(b_\theta^o)] + F_\omega^F(b_\theta^o)[1 - F_\omega(b_\theta^o)]} \tag{53}$$

Therefore, we derive a result consistent with Oren and Sioshansi (2005) that $b_\theta \leq c_\theta$, and the inequality strictly holds when reserve is provided. Recall that efficient profit is given by:

$$\pi_{\theta,\omega}^e = \int_{c_\theta}^{\infty} (p_\xi - c_\theta) dF_\omega \tag{54}$$

Hence, $\pi_{\theta,\omega}^o = \pi_{\theta,\omega}^e$ when $b_\theta = c_\theta$ and $F_\omega^F(p_\xi) = F_\omega(p_\xi)$, which means that all procured quantity is consumed. I summarize the result as follows:

Lemma 4. *In an integrated auction, energy suppliers bid their true cost, while firms that provide reserves have an incentive to shade their bids. In the long run, the investment in flexibility is not efficient.*

Proof. Long-run equilibrium forward price is p_ω^F, so the expected profit of a firm indexed by c is given by:

$$\pi_{\theta,\omega} = (p_\omega^F - b_\theta) + (b_\theta - c_\theta)(1 - F_\omega(b_\theta)) \tag{55}$$

if $b_\theta < p_\omega^F$, and zero otherwise. Efficient bidding is:

$$b_\theta^e = c_\theta - \frac{\int_{-\infty}^{c_\theta} (c_\theta - p_\xi) dF_\omega}{F_\omega(b_\theta^e)} \neq b_\theta^o \tag{56}$$

□

First and foremost, we argue that opportunity cost is not a proper benchmark to reimburse reserve provision, as they should also earn a flexibility premium. Hence, a truthful bidding must lead to under-investment in flexibility, while bid shading that allows reserve providers to earn information rent might be closer to the social optimum. However, the information rent and flexibility premium are not equivalent in a general case, since winners in the auction are paid at least capacity cost for sure, which distorts their incentive in bidding compared to a real-time market where firms have to self-commit and payment is totally uncertain.

4.4.2 Separate Reserves Market: Uniform Pricing

In a subsequent reserves market, opportunity cost is also viewed as a benchmark for capacity bidding. Chao and Wilson (2002) propose an incentive-compatible

two-dimensional auction for reserves with uniform settlement for both capacity and energy payment. They argue that the energy payment should be cleared with a real-time price, and the capacity bid should be the difference between the foregone profit in the forward energy price and the expected profit from the called energy paid at the real-time price. Therefore, it is not surprising that the optimal bid is negative in the example of that paper. We prove that a negative bid is not a special case, but an inevitable result of the underestimation of foregone profit. In theory, if energy provision is cleared by real-time price, the capacity bid should be equal to zero in the absence of risk aversion and price cap, so there is no point in having a capacity market for reserves.

When capacity and activation price are both predetermined, uniform pricing for reserves cannot achieve the long-term optimum, even if reserve suppliers have realized that they should earn the flexibility premium. For technology indexed by θ and provides reserves in state ω, the efficient expected profit should be:

$$\pi_{\theta,\omega} = p^F_\omega + \nu_{\theta,\omega} - c_\theta$$

We propose two separate auctions for upward and downward reserves. Denote $p^{K,\{U,D\}}$ the capacity payment and $p^{X,\{U,D\}}$ the strike price, for upward and downward reserves separately. If the equilibrium uniform price pair $(p^{K,\{U,D\}}, p^{X,\{U,D\}})$ exists, it should satisfy

$$p^{K,U}_\omega - I^U_\theta + [1 - F_\omega(c^U_\theta)][p^{X,U}_\omega - c^U_\theta] = \pi_{\theta,\omega}, \quad \forall\theta \tag{57}$$

for upward reserves, and

$$p^{K,D}_\omega - I^D_\theta + F_\omega(c^D_\theta)[p^{X,D}_\omega + c^D_\theta] = \nu_{\theta,\omega}, \quad \forall\theta \tag{58}$$

for downward reserves. Hence, the system of linear equations in (57) or (58) has a unique set of solutions if and only if two technologies providing reserves, and these two constraints are consistent. There is no solution when more than two types of reserve providers are available, and multiple equilibria exist if there is only one flexible technology.

Nevertheless, in a two-stage uniform auction, competition among firms would imply truthful bidding, meaning they only consider foregone profit in the forward energy market and direct costs to provide reserves, if any, as a flexibility premium behaves as a reward instead of a cost. Since the quantity of required reserves is normally set inelastic, there is no scarcity rent. Therefore, the last technology

that provides flexibility does not earn a flexibility premium if the system operator does not pay an extra payment, and a recursive inference shows that no flexible assets exist in the long run.

Lemma 5. *In a two-stage reserves auction with uniform pricing, each bidder bids its true marginal cost $b_\theta^X = c_\theta^U$ in the second stage, and the difference of foregone profit in the forward energy market and expected energy payment $\mathbb{E}[\pi_\omega^E(\theta)]$ plus commitment cost as capacity bid $b_\theta^K = p_\omega^F - c_\theta + I_\theta^U - \mathbb{E}[\pi_\omega^E(\theta)]$. However, in the long run, no flexibility is provided.*

4.4.3 Separate Reserves Market: Pay-as-bid

Finally, let's check the viability of a pay-as-bid scoring auction. Oren and Bushnell (1994) have shown that for a two-dimensional bid (b^K, b^X), where b^K is the capacity payment for standby and b^X is the price for employed reserve, truthful report of b^X requires a nonlinear scoring rule

$$\mathbb{S}_\omega = b^K + \int_0^{b^X} [1 - F_\omega(p_\xi)] dp_\xi \tag{59}$$

and suppliers agree with the system operator on the probability distribution of energy calls $F_\omega(p)$. However, there must be a markup for the capacity bid. Since there is no profit from activation, the capacity cost should be the foregone profit in the forward energy market on top of the associated commitment cost $p^F - c + I^U$. The objective of a risk-neutral bidder indexed by θ is to maximize

$$\mathbb{E}(\pi(b_\theta^K, b_\theta^X)|\theta) = [b_\theta^K - (p^F - c_\theta + I_\theta^U)]\mathbb{P}_\omega(\mathbb{S}_\omega(b_\theta^K, b_\theta^X), \mathbb{S}_{-i}) \tag{60}$$

$\mathbb{P}$ is the probability that a reserve provider is selected. The optimal bidding is expressed by

$$b_{\theta,\omega}^{K,o} = -\frac{1}{\ln \mathbb{P}_\omega[\mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta^U)]\prime} + p^F - c_\theta + I_\theta^U \tag{61}$$

where

$$\ln \mathbb{P}_\omega[\mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta)]\prime = -\frac{\partial \mathbb{P}_\omega[\mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta)]/\partial \mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta)}{\mathbb{P}_\omega[\mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta)]} \tag{62}$$

A direct result is $b_{\theta,\omega}^{K,o} \geq p^F - c_\theta + I_\theta^U$, if $\frac{\partial \mathbb{P}}{\partial \mathbb{S}} \leq 0$ and $\mathbb{P} \geq 0$. The inequality holds for sure as long as the capacity component affects the score and probability of winning the auction. The intuition is that the standby cost only affects the probability of being selected in the auction, but will not influence the *order of merit*. Therefore, reserve providers are able to earn information rent by bidding

over the short-term capacity cost. To equalize information rent and flexibility premium, the probability of being selected $\mathbb{P}_\omega$ is given by

$$\mathbb{P}_\omega(\mathbb{S}_\omega) = e^{-\int_0^{\mathbb{S}_\omega} \frac{1}{V_\omega(t)} dt} \tag{63}$$

where $\nu_\omega = V_\omega(\mathbb{S})$. One necessary assumption is that the flexibility premium is a function of score, so we rule out the possibility that flexible firms with different flexibility values have the same score in equilibrium. However, by Lemma **??**, to impose information rent to be equal to flexibility premium gives the result $\frac{d\mathbb{S}}{d\theta} = 0$, meaning that all types have the same score in equilibrium, which is contradictory to technology-specific rent. The main intuition is that the change of capacity or energy cost offsets the change of rent: a higher production cost implies a lower capacity cost but higher rent, while a higher flexibility investment (adjustment) cost implies a higher capacity (energy) cost but a lower rent. When the rent is equal to the flexibility premium, the effects are perfectly canceled out. Hence, for each state ω, at most one type of reserve provider is able to earn efficient rent.

Lemma 6. *In a pay-as-bid scoring auction with scoring rule $\mathbb{S}_\omega$ is given by Eq.(59), each bidder bids its true marginal cost $b^X_\theta(\theta) = c^U_\theta$ for energy bid, and the difference between foregone profit in the forward energy market plus information rent as capacity bid, which is given by Eq.(61). However, in the long run, at most one type of flexible asset survives.*

In short, a well-functioning real-time market is the only market-based approach to achieve efficient short-term pricing and long-term investment in flexibility. Reserve markets that centrally provide procurement menus can be a good compromise if real-time settlement is impossible. However, this market is demanding in design as it depends on the time-varying data specified by the system operator. It is, hence, highly sensitive to the information collected and its computation capability. Without extra reimbursement from the system operator, none of the market-based auctions for reserves can restore investment efficiency because they fail to price in technology-specific flexibility premiums. It is pivotal that each generator chooses among different levels of capacity payments in exchange for being available in real-time at the corresponding activation price. Technology-specific flexibility premium guarantees the selection is incentive-compatible. The comparison of different market designs is summarized below:

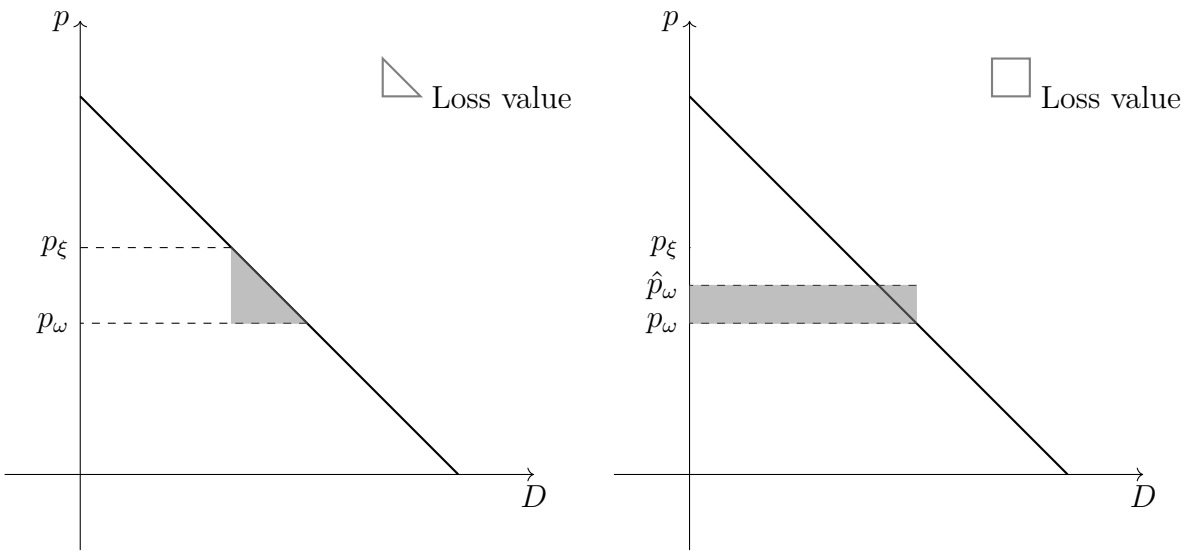
Proposition 5. *The social welfare of each market design is ordered such that:*

$$W(RT) = W(menu) \geq W(scoring) > W(Integrated) = W(uniform) = W(DA\ only) \tag{64}$$

4.5 Demand Flexibility

Rationing The results in (b) and (a) are the same as in Joskow and Tirole (2007), showing that consumers who can react to real-time prices are never rationed, and the rationing for price-insensitive consumers are settled such that the value of lost load is equal to or smaller than real-time price. On top of that, this paper explicitly gives the conditions under which rationing happens for price-insensitive consumers.

Lemma 7. *(1) When rationing can be implemented efficiently, rationing happens when $p_\xi > p_\omega$, and $0 \leq \alpha_\xi < 1$, the value of lost load is equal to real time price: $VOLL_\xi = p_\xi$. (2) When rationing can only be implemented randomly, $VOLL_\xi$ is given by the average surplus of consumers who should have been served without rationing: $VOLL_\xi = \frac{S_\varepsilon(p_\omega)}{D_\varepsilon(p_\omega)} = \hat{p}_\xi \geq p_\omega$; rationing happens when real time price is larger than $VOLL_x i$: $p_\xi > \hat{p}_\xi \Leftrightarrow \alpha_\xi = 0$.*



(a) Effective Rationing **(b)** Random Rationing

Figure 4: Rationing for Price-insensitive Consumers

Share of inflexible consumers The first-order conditions in Proposition 1(b) are independent of the proportion of price-insensitive consumers σ. However, the

share of consumers affects total demand, which in turn has an impact on price sequence $\{F_\omega(p_\xi)\}$ and the net value of alternatives.

Proposition 6. *(1) If rationing can be done efficiently, when the share of price-sensitive consumers increases, production commitment becomes more attractive; (2) If rationing is random, production commitment becomes more attractive in state ω when there is large probability of extreme cases (high demand & low demand), while flexibility is more valuable when moderate cases are more likely to happen.*

Proposition 6 is immediately proven by lemma 3 and 7. When priority service is possible, $VOLL_\xi = p_\xi$, the demand of price-insensitive consumers does not change when they are rationed or they react to real-time price. However, in the no-rationing region, the total demand increases when price-insensitive consumers become price-sensitive since $p_\omega > p_\xi$. Hence, the real-time price would increase, and this decreases the flexibility premium.

However, when rationing can only be done randomly, demand and real-time prices increase in the rationing region ($\hat{p}_\xi < p_\xi$), and in the region where the real-time price is smaller than the marginal price ($p_\xi < p_\omega$), while it decreases in the region where there is no rationing but the real-time price is larger than marginal price ($p_\omega < p_\xi < \hat{p}_\xi$). Therefore, on average, the change of flexibility value is ambiguous.

The intuition of this result is: when total demand is more elastic, there is less risk of curtailment, so less cost of production commitment (curtailment effect). But this elasticity also implies less demand when the price is moderately high (price effect). Therefore, there are two countervailing effects, and for those technologies with low curtailment cost but high production cost, price effects can dominate, and production commitment becomes less attractive when there are more price-sensitive consumers. Hence, demand flexibility does not necessarily reduce the requirement for production flexibility.

5 Conclusion

This article analyzes efficient pricing and investment of (in)flexible technologies, when products are hard to store and demand is uncertain. There are three take-aways. First, efficient pricing is state-contingent, reflecting the flexibility premium and inflexibility costs: flexible firms can earn more than the expected price, and inflexible firms produce even if short-run profit is negative. Second, in the absence

of a real-time market, a forward-only market would result in investment distortion. Finally, a reserves market with a menu of call options, combined with a forward market, can achieve the social optimum as in the real-time market.

This model is relevant to electricity markets. It indicates the importance of a real-time or balancing market to provide correct incentives for flexibility investment, other than a complementary market to adjust any imbalance. Moreover, it implies the drawbacks of the main proposed reserves market design. Under general conditions, scoring auction, co-optimization, or predetermined uniform pricing will lead to flexibility under-investment, as they fail to reimburse the technology-specific flexibility premium. Nowadays, most European countries develop both a balancing energy market and a capacity market, arguing that flexibility providers need to reimburse their startup and adjustment costs via the capacity market, which is not supported by this paper. We show that balancing price can reimburse any direct and indirect costs associated with the adjustment. Therefore, when there is no risk-aversion and other market failures, a balancing capacity market is not necessary to complement the balancing energy market.[18]

Moreover, this paper builds an asymmetric structure of inflexibility cost and introduces investment cost in flexibility. This complexity is not unnecessary, but observed from practice. One example is a wind farm, which is easy to turn off but not to turn on: its generation also depends on the weather, which cannot be controlled. This assumption motivates the separate purchase of incremental and decremental reserves. Furthermore, the investment cost in flexibility is also reasonable, as plants need a warm-up to produce. By this assumption, we not only have a rough idea about adjustment costs, but also how they explicitly enter the cost function and equilibrium. This cost structure is also welcomed to be tested by empirical research.

Last, even though this paper is in the context of the electricity market, a flexibility premium widely exists in other sectors, such as transportation and hotels: consumers choose to book non-refundable tickets or hotels or a refundable one with a flexible premium, or buy a ticket at the last minute with a more volatile price. Airlines provide flexibility by purchasing different sizes of airplanes and hiring more people as reserves. Our conclusion can also be applied to these fields.

[18] In practice, risk aversion, market power, and price cap can rationalize the capacity market, but reimbursement for direct adjustment payment is not a good reason.

References

Anderson, E., Chen, B., & Shao, L. (2017). Supplier Competition with Option Contracts for Discrete Blocks of Capacity. Operations Research, 65(4), 952-967.

Anupindi, R., & Jiang, L. (2008). Capacity Investment Under Postponement Strategies, Market Competition, and Demand Uncertainty. Management Science, 54(11), 1876-1890.

Boiteux, M (1960). Peak-load Pricing. The Journal of Business, 33(2), 157-179.

Borenstein, S. & Holland S. P. (2005). On the Efficiency of Competitive Electricity Markets With Time-Invariant Retail Prices. The RAND Journal of Economics, 36(3), 469-493.

Bushnell, J., & Oren, S. (1994). Bidder Cost Revelation in Electric Power Auctions. Journal of Regulatory Economics, 6, 5-26.

Brown, G. & Johnson, M.B. (1969). Public Utility Pricing and Output under Risk. American Economic Review, 59(1), 119-128.

Brijs, T., De Jonghe, C., Hobbs, B.F., & Belmans, R. (2017). Interactions between the Design of Short-term Electricity Markets in the CWE Region and Power System Flexibility. Applied Energy, 195, 36-51.

Carlton D.W. (1977). Peak Load Pricing with Stochastic Demand. The American Economic Review, 67(5), 1006-1010.

Crew, M.A. & Kleindorfer, P.R. (1976). Peak Load Pricing with a Diverse Technology. Bell Journal of Economics, 7(1), 207-231.

Chao, HP. & Wilson, R. (1987). Priority Service: Pricing, Investment and Market Organization. The American Economic Review, 77(5), 899-916.

Chao, HP. & Wilson, R. (2002). Multi-Dimensional Procurement Auctions for Power Reserves: Robust Incentive-Compatible Scoring and Settlement Rules. Journal of Regulatory Economics, 22, 161-183.

Cramton, P. (2017). Electricity Market Design. Oxford Review of Economic Policy, 33(4), 589-612.

Denholm, P., & Hand, M. (2011). Grid Flexibility and Storage Required to Achieve Very High Penetration of Variable Renewable Electricity. Energy Policy, 39(3), 1817-1830.

Gould, J.P. (1968). Adjustment Costs in the Theory of Investment of the Firm. The Review of Economic Studies, 35(1), 47-55.

Hay, G.A. (1970). Adjustment Costs and the Flexible Accelerator. The Quarterly Journal of Economics, 84(1), 140-143.

Hogan, W. W. (2013). Electricity Scarcity Pricing Through Operating Reserves. Economics of Energy & Environmental Policy, 2(2), 65-86.

Hogan, W. W. (2005). On an "Energy Only" Electricity Market Design for Resource Adequacy. Working Paper. Retrieved from https://scholar.harvard.edu/whogan/files/hogan_energy_only_092305.pdf.

Hortaçsu, A., & Puller, S.L. (2008). Understanding Strategic Bidding in Multi-Unit Auctions: A Case Study of the Texas Electricity Spot Market. RAND Journal of Economics, 39(1), 86-114.

Joskow, P. & Tirole, J. (2007). Reliability and Competitive Electricity Market. RAND Journal of Economics, 38(1), 60–84.

Ito, K., & Reguant, M. (2016). Sequential Markets, Market Power, and Arbitrage. American Economic Review, 106(7), 1921-57.

Jaramillo, F., Schiantarelli, F., & Sembenelli, A. (1993). Are Adjustment Costs for Labor Asymmetric? An econometric Test on Panel Data for Italy. The Review of Economics and Statistics, 75(4), 640-648.

Kleindorfer P.R., & Wu D.J. (2005). Competitive options, supply contracting, and electronic markets. Management Science, 51(3), 452-466.

Kondziella, H., & Bruckner, H. (2016). Flexibility Requirements of Renewable Energy Based Electricity Systems – a Review of Research Results and Methodologies. Renewable and Sustainable Energy Reviews, 53, 10-22.

Lucas, R.E. (1967). Adjustment Costs and the Theory of Supply. Journal of Political Economy, 75(4), 321-334.

Lund, P.D., Lindgren, J., Mikkola, J., & Salpakari, J. (2015). Review of Energy System Flexibility Measures to Enable High Levels of Variable Renewable Electricity. Renewable and Sustainable Energy Reviews, 45, 785-807.

Oren, S.S. (2003). Ensuring Generation Adequacy in Competitive Electricity Markets. UC Berkeley: University of California Energy Institute. Retrieved from https://escholarship.org/uc/item/8tq6z6t0.

Oren, S.S., & Sioshansi, R (2005). Joint Energy and Reserves Auction with Opportunity Cost Payment for Reserves. International Energy Journal, 6(1), (4)35-(4)44.

Pfann, G.A., & Verspagen, B (1989). The Structure of Adjustment Costs for Labor in the Dutch Manufacturing Sector. Economics Letters, 29(4), 365-371.

Pfann, G.A., & Palm, F. C. (1993). Asymmetric Adjustment Costs in Labor Demand Models with Empirical Evidence for the Dutch and UK Manufacturing Sectors. The Review of Economic Studies, 60(2), 397-412.

Puera, H & Bunn, D. W. (2022). Renewable Power and Electricity Prices: The

Impact of Forward Markets. Management Science, 67(8), 4772-4788.

Schramm, R. (1970). The Influence of Relative Prices, Production Conditions and Adjustment Costs on Investment Behaviour. The Review of Economic Studies, 37(3), 361-376.

Sedzro, K.S.A., Kishore, S., Lamadrid, A. J., & Zuluaga, Luis F. (2018). Stochastic Risk-sensitive Market Integration for Renewable Energy: Application to ocean wave power plants. Applied Energy, Elsevier, 22, 474-481.

Schwartz, E.S., & Trigeorgis, L. (2004). Real options and investment under uncertainty: classical readings and recent contributions (1st ed.). MIT Press.

Trigeorgis, L. (1996). Real options: Managerial flexibility and strategy in resource allocation. MIT Press.

Visscher, L.M. (1973). Welfare-maximizing Price and Output with stochastic Demand. The American Economic Review, 63(1), 224-229.

Van Der Weijde, A.H., & Hobbs, B.F. (2012). The Economics of Planning Electricity Transmission to Accommodate Renewables: Using Two-stage Optimisation to Evaluate Flexibility and the Cost of Disregarding Uncertainty. Energy Economics, 34(6), 2089-2101.

Wilson, R. (2002). Architecture of Power Plants. Econometrica, 70(4), 1299-1340.

Appendix

A Proof of Proposition 1 (d)

Proof. Take derivative w.r.t $Q_{\theta,\omega}, R^D_{\theta,\omega}$ and $R^U_{\theta,\omega}$:

$$[Q_{\theta,\omega}] : \mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta - \lambda_{\theta,\omega} + \mu_{\theta,\omega} + \varphi^Q_{\theta,\omega} = 0 \tag{A.1}$$

$$[R^D_{\theta,\omega}] : \mathbb{E}_{\xi|\omega}[u^D_{\xi,\theta}(c^D_\theta - p_\xi)] - I^D_\theta - \mu_{\theta,\omega} + \varphi^D_{\theta,\omega} = 0 \tag{A.2}$$

$$[R^U_{\theta,\omega}] : \mathbb{E}_{\xi|\omega}[u^U_{\xi,\theta}(p_\xi - c^U_\theta)] - I^U_\theta - \lambda_{\theta,\omega} + \varphi^U_{\theta,\omega} = 0 \tag{A.3}$$

(A.2) and (A.3) can be rewritten as

$$\mathbb{E}_{\xi|\omega}[\max\{c^D_\theta - p_\xi, 0\}] - I^D_\theta - \mu_\omega + \varphi^D_{\theta,\omega} = 0 \tag{A.4}$$

$$\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c^U_\theta, 0\}] - I^U_\theta - \lambda_\omega + \varphi^U_{\theta,\omega} = 0 \tag{A.5}$$

which correspond to a put option and a call option respectively. If a firm does not invest in downward flexibility, $dR^D_{\theta,\omega} < dQ_{\theta,\omega} \Leftrightarrow \mu_{\theta,\omega} = 0$, (A.1) is reduced to a forward:

$$\mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta - \lambda_{\theta,\omega} = 0 \tag{A.6}$$

If a firm invests in downward flexibility, both (A.1) and (A.4) hold, and to sum up these two equations gives

$$\begin{aligned}
&\mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta + \mathbb{E}_{\xi|\omega}[\max\{c_\theta^D - p_\xi, 0\}] - I_\theta^D - \lambda_{\theta,\omega} = 0 \\
\Rightarrow &\mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta - \mathbb{E}_{\xi|\omega}[\min\{p_\xi - c_\theta^D, 0\}] - I_\theta^D - \lambda_{\theta,\omega} = 0 \\
\Rightarrow &\mathbb{E}_{\xi|\omega}[p_\xi - c_\theta^D] - c_\theta - \mathbb{E}_{\xi|\omega}[\min\{p_\xi - c_\theta^D, 0\}] + c_\theta^D - I_\theta^D - \lambda_{\theta,\omega} = 0 \\
\Rightarrow &\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta^D, 0\}] - [c_\theta - c_\theta^D] - I_\theta^D - \lambda_{\theta,\omega} = 0
\end{aligned}$$

Therefore, a combination of a forward and a put option is equivalent to a call option. □

B Proof of Proposition 4

Proof. The procurement and activation strategies of the retailers are straightforward, and we focus on the self-selection of the generators. For any θ choosing upward reserves, incentive compatibility requires that $\forall \theta, X$,

$$\mathbb{E}_{\xi|\omega}\{\max(\hat{p}_\xi - c_\theta^{C/D}, 0)\} \geq \mathbb{E}_{\xi|\omega}\{\max(\hat{p}_\xi - X, 0)\} + [1 - F_\omega(X)][X - c_\theta^{C/D}] \quad \text{(B.1)}$$

This is equivalent to

$$\int_{c_\theta^{C/D}}^{X} \hat{p}_\xi dF_\omega \geq [F_\omega(X) - F_\omega(c_\theta^{C/D})]c_\theta^{C/D} \quad \text{(B.2)}$$

which always holds. □