Flexibility in Power System: Market Design Matters

Dongchen He*, Bert Willems†

September 24, 2024

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, day-ahead forward price results in under-investment in flexible technologies and over-investment in inflexible ones. This distortion, in theory, can be corrected by a time-varying options market with technology-specific payment while any centralized auction fails to achieve 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 decision when quick adjustment is expensive. This question is particularly pertinent to electricity markets, demand of which is so unpredictable while to continuously balance the market is vital to social and economic development; the energy transition further raises the concern about the electricity supply reliability. According to the report "Net Zero by 2050" from the International Energy Association (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 are from hydro-power.

With an ambitious investment in global intermittent renewable energy sources (IRES) over the next three decades, IEA predicts a four-fold 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 incentivise 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, to economically run flexible units is equally important. Flexibility in this context refers to an economic terminology: adjustment cost. Therefore, this paper establishes a model based on adjustment cost and assumes 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 turbine that are rather flexible.

Retrofits or investments of new flexible units incur a higher investment cost

[1] Flexibility requires more than double by 2030 and 7 times by 2050 in EU, according to 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 purchase 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 independent decision for each state.

in exchange of reducing adjustment cost. For example, the hydrogen to power system is very flexible while expensive. Solar panel with battery is flexible while more costly than standard ones without battery. Coal-fired power plants can improve its operational flexibility by retrofitting with 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, 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 supply of flexibility, providing simple numerical results. With an abundance of economics 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 the time when renewable sources are negligible. Reserve markets are designed to induce short-term flexibility but efficiency in long term is not well addressed. This work aims at filling 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 design(s) is(are) able to achieve this efficiency?*

The results show 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, day-ahead forward price takes the incentive to invest in flexible technologies while induces over-investment in inflexible ones. Complementing with a reserve market [3] in which the system operator (SO) provides an array of time-varying and technology-specific contracts can restore optimum. Nevertheless, the existing reserve market design such as a day-ahead 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 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 model, our supply function is steeper and 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 day-ahead 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 who sell reserves should earn a technology-specific flexibility premium on top of the opportunity cost of not trading in the day-ahead energy market and adjustment costs related. The flexibility value is called option premium or value of waiting [4] in real options theory (Trigeorgis, 1996; Schwartz and Trigeorgis, 2004), which is the difference between real options value and net present value. Reserves are not only 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 flexibility investment .

Note that social optimum is hardly obtained by a combination of market-based day-ahead 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

[3] In the context of this paper, a day-ahead market is a forward market, and a reserve market plays the role of an option market.

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

words, in the absence of risk aversion and market power, when efficient real-time pricing is available, a separate reserve market is unnecessary. [5]

In practice, it is difficult that all transactions are settled nearly instantaneously. When approaching real-time, things become more certain, so a day-ahead 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 to 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 that asserts the significance of real-time market to compensate for flexibility, but there is no way to repudiate 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 reason 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 of demand response does not necessarily reduce the need for production flexibility.

The rest of paper is organized as follows. Section 2 provides an overview of the relevant research. Section 3 constructs the model. Section 4 illustrates the impact of different market design. Sections 5 gives a simple example and section 6 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

[5] As 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 at least to reserve capacity that is able to keep grid stable in the event of an unexpected outage of the largest 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.

the development of literature and highlight the contribution of this paper.

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.

Modelling stateic but deterministic demand supplied by a single technology, Boiteux (1960)[7]shows the offpeak 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 both offpeak 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 stateic, but also uncertain (Visscher, 1973; Carlton, 1977; Brown and Johnson, 1969, Crew and Kleindorfer 1976). The problem becomes to choose 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 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 additive term, price is lower than long-run marginal cost, but the conclusion is reverse with multiplicative demand uncertainty.

The theoretical progress stops 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 of free, and ignoring this cost would lead to inefficient investment. Second, even 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 decision and value of commitment by saving adjustment costs, which in turn affects capacity decision.

This paper extends the classic two-stage peak-load pricing model to three stages of which production commitment is in between investment and production, arguing that efficient real-time prices should reflect cost of adjustment and

[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.

the value of flexibility. We model both stateic and uncertain demand net of renewables, 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 a 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 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 short term, and adjustments to demand are staggered. Therefore, long-run equilibrium is not the minimal point of a U-shape cost function, but also includes costs to approaching and keeping it.

Adjustment costs are analyzed in econometric studies in labor demand, and the core discussion is the costs' structure: whether hiring or firing costs are symmetric? While the standard assumption to adjustment cost is a symmetric and quadratic function, data collected from Italy (Schramm, 1970), the Netherlands and the UK (Pfann and Verspagen, 1989; Pfann and Palm, 1993) rejects this hypothesis. There is also a strand of energy economics literature considering inflexibility in real-time power market by introducing a steeper supply curve in real-time compared to a day-ahead market (Ito and Reguant, 2016; Hortaçsu and 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 adjustments costs and how they are reflected in equilibrium behavior. Dispatchable generators are an indicator of supply side electricity flexibility, and what economists should do is to 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 deviation incurs 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 blackout (Cramton, 2017). Hence, reserves are usually regarded as a way to provide reliability (Bushnell and 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 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, to provide reserves is necessary to avoid system collapse. Reserves playing a role of providing additional capacity and avoiding possibly huge loss is a natural and standard reliability consideration.

On top of that, reservew are also viewed as financial hedge [9] to deal with spot price uncertainty. In two related papers (Kleindorfer and Wu (2005), Anderson et al. (2017)), reserves are used as an options contract. Instead of a "backup" role, reserves are strategically substitutes of 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 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 capacity and production part of reserves and to deal with asymmetrical information of 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 spot

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

[10]The merit order is a way of ranking dispatch of electricity, based on ascending order of pricewhich 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.

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 and Oren (1994)). Oren and Sioshansi (2005) propose an integrated market for both energy and reserves in which called reserves are paid the same price as energy and reserves not activated receive a capacity payment equal to the difference between 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 a technology-specific flexibility value in reserve payment. This finding shows that none of the uniform pricing auction proposed by Chao and Wilson (2002), integrated auction by Oren and Sioshansi (2005), and the nonlinear scoring auction by Bushnell and Oren (1994) can guarantee proper level of flexibility.

3 Model

Electricity, generated by multiple technologies, cannot be cost-effectively stored.[11] Its demand is uncertain while satisfying stochastic demand is important for grid stability and social welfare. Supply uncertainty[12] from renewables further raises concerns about flexibility of continuously balancing supply and demand.

Three categories of agents are active in electricity markets: generators, retailers, and consumers. All agents are risk-neutral.[13] Generators and retailers trade in the wholesale market and retailers sign contract with consumers. For the sake of simplicity, there is no cost for retail activity. If the wholesale market is real-time and all consumers can react to real-time prices, retailers and consumers are equivalent from the model perspective. Otherwise, the absence of real-time markets or failing to react to real-time prices rationalizes the existence of retailers

[11] Strictly speaking, electricity cannot itself be stored on any scale, but it can be converted to other forms of energy such as mechanical energy, thermal energy or chemical energy, and be converted back when needed. However, the investment of storage capacity is restrictive and expensive, while the conversion is not so efficient. Hence, electricity storage has not been widely adopted and we omit it here.

[12] Demand in this paper is defined as the total demand minus renewable generation. Hence, supply uncertainty has been internalized in net energy demand distribution.

[13] We simplify the risk preference to isolate the effects of (in)flexibility. A different risk preference, even alters some results, does not change the main conclusion in this paper.

to implement a two-part tariff.[14] Inflexibility comes from two sources: supply and demand. Supply inflexibility means adapting output is costly on short notice. Demand inflexibility reveals the fact that consumers are not sensitive to real-time prices so real-time demand is rather inelastic.

Generators provide power in two potential ways: (1) *energy*, (2) *reserves*. Energy is the initial amount a firm commits to produce, and reserves are available capacity not currently used but can quickly adapt output to unexpected shocks in real-time. In electricity markets, a forward wholesale market selling energy is called the day-ahead market and the reserve market is an options market. Reserves are divided into two types[15]: (1) *upward reserves* are the backup to increase generation, and (2) *downward reserves* are activated to curtail production.

Generators proceed with a three-stage game. In the first stage (investment), they choose which technology(ies) and how much capacity to invest. Then (day-ahead commitment), generators make production and reserve commitments under capacity constraints using partially revealed demand information. In the last stage (real-time realization), actual demand is known; generators produce and/or activate reserves under commitment constraint.

3.1 Uncertainty

In the second stage, nature draws information set ω, frequency of which follows the density distribution $g(\omega)$. In our context, ω consists of information such as the hour of the next day, the anticipated demand and weather condition as well as their distributions $\phi(\varepsilon|\omega)$[16]. For example, today, firms have expectation of demand and wind/solar supply for four o'clock tomorrow afternoon. In the last stage, demand (net of renewables) ε is revealed. Define $\xi = (\omega, \varepsilon)$, and the joint density function of ξ is

$$h(\xi) = g(\omega) \cdot \phi(\varepsilon|\omega)$$

In the first stage, investment decision is made by taking expectations over ξ. In the second stage, by knowing ω, firms make commitment decision according to $\phi(\varepsilon|\omega)$, and real production quantity is determined based on revealed ε in the

[14] We briefly explain it in Section 5 with a simple example.

[15] There are different ways to categorize reserves. For example, according to the activation time and the duration of activation, reserve power is divided into primary reserve, secondary reserve and tertiary reserve. Also, depending on whether generating capacity currently connected to the system or not, reserve can be divided into spinning reserve and non-spinning reserve. In our paper, we only categorize them by the direction of adjustment.

[16] Electricity prices have evident daily, weekly and seasonal effects. Hence, we assume state-contingent density function.

final round.

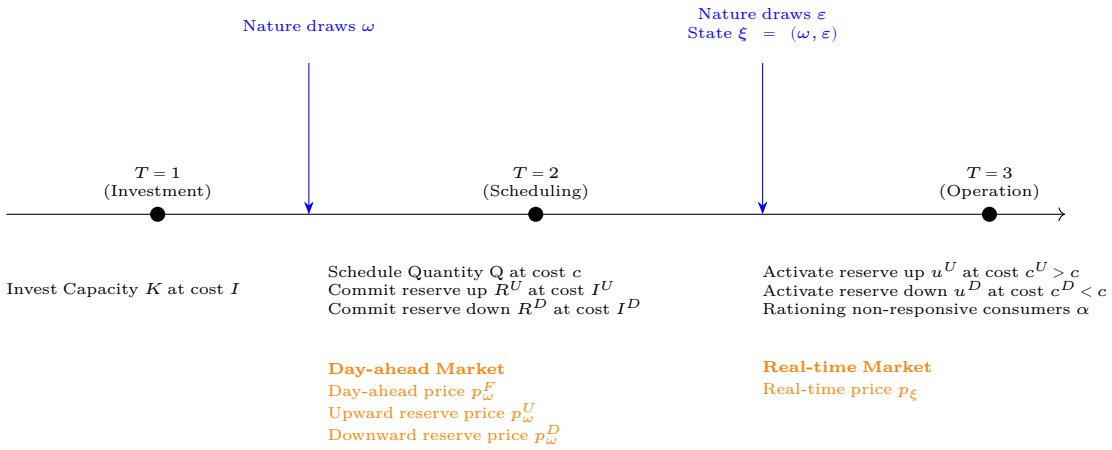


Figure 1: Timing of Decision and Information Revelation

3.2 Supply Side

To illustrate the flexibility characteristics of generating technology, we start with a single technology scenario, followed by a more general case incorporating multiple technologies.

3.2.1 Single Technology

A technology is described by six cost parameters which represent investment, flexibility and production costs. See table 1.

c	marginal production cost without adjustment
c^U	marginal production cost for upward adjustment
c^D	marginal production cost for downward adjustment
I	per unit investment cost
I^U	per unit commitment cost for upward reserve
I^D	per unit commitment cost for downward reserve

Table 1: Cost Parameters for Technology

To be clear, c is the production cost if there is no deviation from original generation schedule, and I is the long-term capacity investment cost. I^U and I^D are flexibility commitment costs incurred in the second stage to be able to adjust output in the final round. This is a new attribute of this paper, reflecting the fact that power outage can happen even though there is extra installed capacity.

These are direct costs of providing reserves such as startup and no-load cost.[17] Without investing I^U or I^D, the supply function in the short-run is vertical. c^U is the production cost of called upward reserve. c^D is the marginal cost of executed downward reserve.

The total production q_ξ in state ξ is given by the production Q_ω scheduled at state ω, plus a term for the fraction u_ξ^D of upward reserves R_ω^U that are activated in real time, minus a term the fraction u_ξ^D of downward reserves R^D that activated in real time:

$$q_\xi = Q_\omega + u_\xi^U R_\omega^U - u_\xi^D R_\omega^D \tag{1}$$

The sum of scheduled production and upward reserves needs to be less than installed capacity and the amount of downward reserves needs to be less than committed production capacity. So scheduling decisions need to satisfy

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

Ex-post the cost incurred in state ξ is given by the investment cost IK, the cost of scheduling production and upward and downward reserves $cQ_\omega + I^U R_\omega^U + I^D R_\omega^D$, the cost of increasing production in real time $u_\xi^U R_\omega^U c^U$, and the cost savings of reducing production in real rime $u_\xi^D R_\omega^D c^D$.

$$C_\xi = IK + cQ_\omega + (I^U + u_\xi^U c^U) R_\omega^U + (I^D - u_\xi^D c^D) R_\omega^D \tag{3}$$

We assume that it is more efficient to schedule production early, and adjusting production later on comes at a cost: Short-term upward adjustments are assumed to be more costly than planned production, $c^U \geq c$. And with a short-term reduction of output we cannot fully recoup the marginal production costs, as some of those costs are sunk, $c^D \leq c$.

The model has two types of flexibility: upward and downward reserves. In order to avoid trivial solutions where only one type of reserves is used we make additional assumptions. Suppose we need to satisfy demand D with probability P and demand $D+1$ with probability $1-P$. We can achieve this by scheduling a quantity $Q = D$, and one unit of upward reserves $R^U = 1$ or by scheduling a quantity $Q = D+1$ and one unit of downward reserves $R^D = 1$. The cost of

[17] Commitment cost of downward reserves is usually imposed to be zero: $I^D = 0$ (Chao and Wilson (2002)). However, we keep this notation as a general representation. In practice, the costs not only relate to fuels or machine depreciation, but also staff who are employed to investigate the latest change in the market.

satisfying demand for both types of flexibility is given by:

$$\text{Up:} \quad cD \qquad +I^U + c^U(1-P) \tag{4}$$

$$\text{Down:} \quad c(D+1) + I^D - c^D P \tag{5}$$

Using upward reserves to create production flexibility becomes cheaper if we have to activate reserve less often, that is when the probability P of having low demand is large. In that situation, using downward reserves becomes more expensive as we need to activate the reserves more often. We assume that for $P \approx 1$ upward reserves are the cheapest form of flexibility and for $P \approx 0$ downwards reserves are cheaper. This requires that

$$c - c^U < I^U - I^D < c - c^D \tag{6}$$

Figure 2 illustrates the supply curve in the real-time market. If a technology is fully flexible, the commitment stage becomes irrelevant,[18] so the model reduces to a standard two-stage peak-load pricing model with investments and real-time operation. A firm produces state-contingent quantity $q_\xi = Q_\varepsilon$ up to installed capacity K. By contrast, an inflexible technology[19] cannot adjust output in real time and has to commit to production Q_ω according to state ω, and the real-time supply curve is inelastic. Finally, a partially flexible technology is allowed to adjust output at a cost in real time if they invest in flexibility (reserves) R^U_ω, R^D_ω.

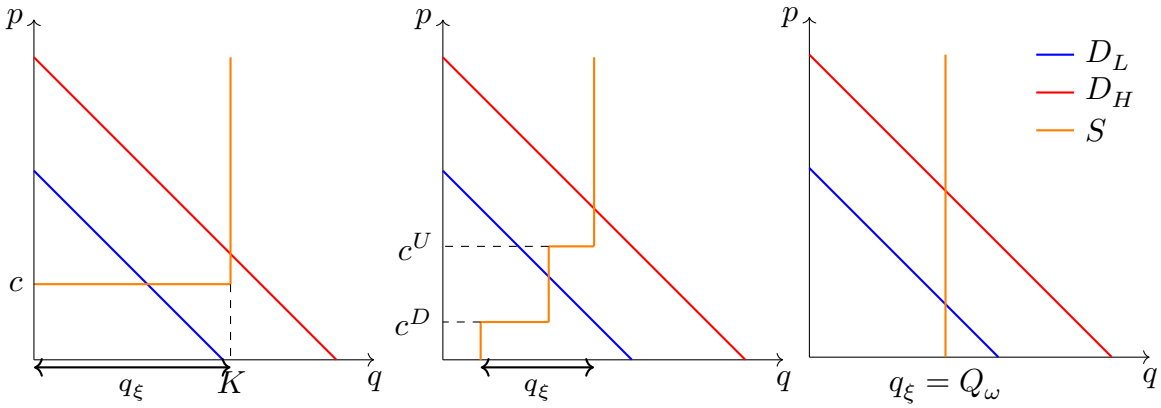
3.2.2 Multiple Technologies

From now on, we proceed to the general scenario encompassing N types of technologies indexed by $\theta \in \{1, 2, ..., N\}$. The total ex-post cost incurred in state ξ is:

$$\begin{aligned} C_\xi &= \sum_{\theta=1}^{N} C_{\theta,\xi} \\ &= \sum_{\theta=1}^{N} I_\theta K_\theta + c_\theta Q_{\theta,\omega} + (I^U_\theta + u^U_{\theta,\xi} c^U_\theta) R^U_{\theta,\omega} + (I^D_\theta - u^D_{\theta,\xi} c^D_\theta) R^D_{\theta,\omega} \end{aligned} \tag{7}$$

[18] $I^U = I^D = 0, c^U = c^D = c$

[19] Inflexibility implies there is no way to adjust in any direction, $I^U = I^D = \infty$, $I^U_\theta = |c^D| = \infty$, $I^D = c^U = \infty$, or $c^U = |c^D| = \infty$



(a) Fully Flexible **(b)** Partially Flexible **(c)** Fully Inflexible

Figure 2: Supply Function of Technologies with Different Flexibility

Note: This figure represents the commitment decision and equilibrium outcome in high demand (D_H) and low demand (D_L) scenarios, when the technology has different level of flexibility. Panel (2a) shows the case when technology is fully flexible; commitment is free of cost so the real-time supply curve (S, in orange) is horizontal across marginal production cost c until reaching capacity constraint K. Panel (2c) shows the case when technology is fully inflexible, in which case adjustment is never possible in real-time, implying actual demand must be equal to quantity commitment $q_\xi = Q_\omega \leq K$. Panel (2b) shows the technology in between, besides quantity commitment Q_ω, the firm can also commit to upward reserve R^U_ω and downward reserve R^D_ω. while the actual demand is restricted between, $q_\xi \in [Q_\omega - R^D_\omega, Q_\omega + R^U_\omega]$.

3.3 Demand Side

Following Borenstein and Holland (2007), consumers are divided into two categories: a total portion of σ price-insensitive consumers and $(1-\sigma)$ price-sensitive consumers. σ is exogenous. Price-sensitive consumers can react to real-time price but price-insensitive consumers are only recorded over the aggregate demand over all states. Except for distinct demand elasticity, there is no other difference between price-insensitive and price-sensitive consumers.

Demand functions in the absence of rationing are $D_\varepsilon(p_\omega)$ and $\hat{D}_\varepsilon(p_\xi)$, for price-insensitive and sensitive consumers respectively, both increasing with ε. Retailers are able to provide the interruptible contract to consumers. Hence, rationing[20] is possible; denote the fraction of satisfied demand by α_ξ and $\hat{\alpha}_\xi$, for price-insensitive consumers and price-sensitive consumers. The expected consumption and gross surplus are

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

[20]Rationing can be efficient or random. Efficient rationing means consumption limiting starts from consumers with lowest willing-to-pay. Random rationing means randomly choosing consumers that will not be served.

for price-insensitive consumers, and

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

for price-sensitive consumers. Surplus function is increasing and concave in consumption. p_{ω} is the marginal price paid by price-insensitive consumers, and p_{ξ} is the real-time price faced with price-sensitive consumers.

The value of lost load (henthforth: VOLL) is defined as the change of marginal surplus associated with a unit increase of supply to consumers:

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

3.4 Social Optimum

The social optimum is 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 $u^{U}_{\theta,\xi}$, $u^{D}_{\theta,\xi}$ and the price p_{ω} for price-insensitive consumers, and their rationing rates α_{ξ}, $\hat{\alpha}_{\xi}$, to maximize the social welfare function:

$$\mathbb{E}_{\xi}\{\sigma \mathcal{S}_{\varepsilon}(p_{\omega}, \alpha_{\xi}) + (1-\sigma)\hat{\mathcal{S}}_{\varepsilon}(p_{\xi}, \hat{\alpha}_{\xi}) - C_{\xi}\} \tag{11}$$

$$\begin{aligned}
\text{s.t.} \quad & \sigma \mathcal{D}_{\varepsilon}(p_{\omega}, \alpha_{\xi}) + (1-\sigma)\hat{\mathcal{D}}_{\varepsilon}(p_{\xi}, \hat{\alpha}_{\xi}) \leq \sum_{\theta=1}^{N} q_{\theta,\xi} && [p_{\xi} h_{\xi}] \\
& Q_{\theta,\omega} + R^{U}_{\theta,\omega} \leq K_{\theta} && [\lambda_{\theta,\omega}] \\
& R^{D}_{\theta,\omega} \leq Q_{\theta,\omega} && [\mu_{\theta,\omega}] \\
& Q_{\theta,\omega} \geq 0 && [\varphi^{Q}_{\theta,\omega}] \\
& R^{D}_{\theta,\omega} \geq 0 && [\varphi^{D}_{\theta,\omega}] \\
& R^{U}_{\theta,\omega} \geq 0 && [\varphi^{U}_{\theta,\omega}]
\end{aligned}$$

Let $p_{\xi} h_{\xi}$ denote the multiplier of production constraint in 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 second best optimum* [21] *of maximization*

[21] The result is first-best if $\sigma = 0$. That is, all consumers can react to real-time prices so there

problem (11) are

(a) Price-insensitive consumers.

$$\begin{aligned} &\forall\omega \qquad \mathbb{E}_{\xi|\omega}\left[\frac{\partial \mathcal{S}_{\varepsilon}}{\partial p_{\omega}} - p_{\xi}\frac{\partial \mathcal{D}_{\varepsilon}}{\partial p_{\omega}}\right] = 0 \\ &\forall\xi \qquad VOLL_{\xi} = \frac{\partial \mathcal{S}_{\varepsilon}}{\partial \alpha_{\xi}} / \frac{\partial \mathcal{D}_{\varepsilon}}{\alpha_{\xi}} = p_{\xi} \quad or \quad \alpha_{\xi} \in \{0,1\} \end{aligned} \tag{12}$$

(b) Price-sensitive consumers.

$$\begin{aligned} &\hat{\alpha}_{\xi} = 1 \\ &\hat{\mathcal{D}}_{\varepsilon} = \hat{D}_{\varepsilon}(p_{\xi}) \end{aligned} \tag{13}$$

(c) Efficient activation of reserves.

$$u^{D}_{\theta,\xi} = \begin{cases} 1, & p_{\xi} < c^{D}_{\theta} \\ \in [0,1], & p_{\xi} = c^{D}_{\theta} \\ 0, & p_{\xi} > c^{D}_{\theta} \end{cases} \qquad u^{U}_{\theta,\xi} = \begin{cases} 1, & p_{\xi} > c^{U}_{\theta} \\ \in [0,1], & p_{\xi} = c^{U}_{\theta} \\ 0, & p_{\xi} < c^{U}_{\theta} \end{cases} \tag{14}$$

(d) Commitment.

$$\begin{aligned} (forward) \quad & \mathbb{E}_{\xi|\omega}[p_{\xi}] - c_{\theta} = \lambda_{\theta,\omega} - \mu_{\theta,\omega} - \varphi^{Q}_{\theta,\omega} \\ (put\ option) \quad & \mathbb{E}_{\xi|\omega}[\max\{c^{D}_{\theta} - p_{\xi}, 0\}] = I^{D}_{\theta} + \mu_{\theta,\omega} - \varphi^{D}_{\theta,\omega} \\ (call\ option) \quad & \mathbb{E}_{\xi|\omega}[\max\{p_{\xi} - c^{U}_{\theta}, 0\}] = I^{U}_{\theta} + \lambda_{\theta,\omega} - \varphi^{U}_{\theta,\omega} \end{aligned} \tag{15}$$

(e) Capacity Investment.

$$\begin{aligned} &\mathbb{E}_{\omega}[\lambda_{\theta,\omega}] = I_{\theta} \quad if \quad K_{\theta} > 0 \\ &otherwise,\ K_{\theta} = 0 \end{aligned} \tag{16}$$

We interpret the first-order conditions as follows:

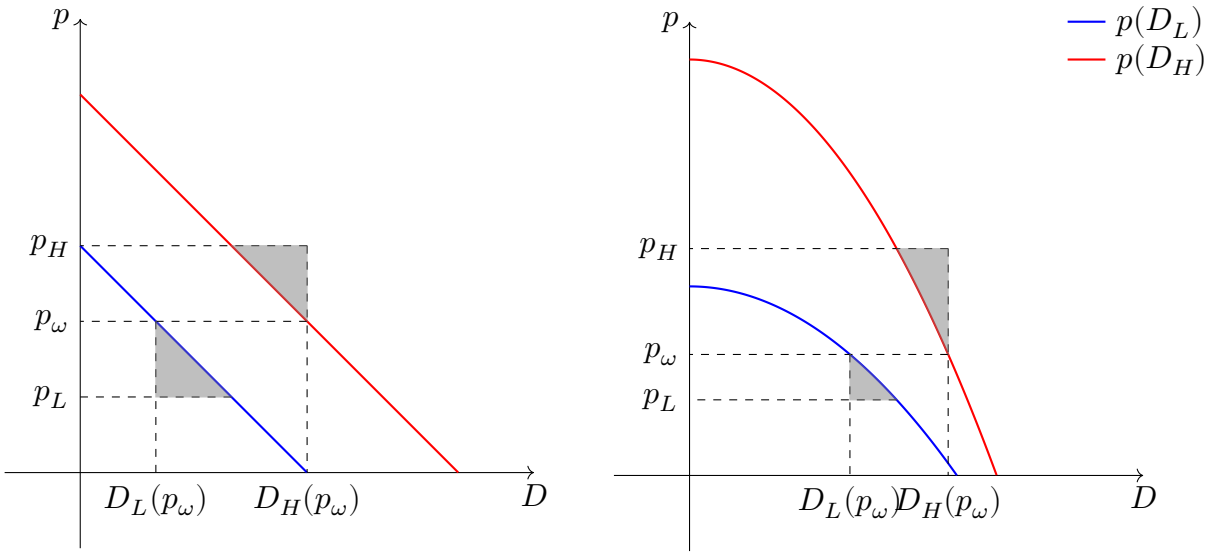
Proposition 1 (a) gives the optimal conditions for price-insensitive consumers. First, in case of rationing, the optimal rationing rate in state ξ should be such that VOLL in that state VOLL_{ξ} is equal to the real-time price p_{ξ}. Note that the real-time price indicates the shadow price of production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

is no welfare loss from rationing.

the social value of energy in state ξ. The price that consumers pay in state ω, p_ω is such that the losses to distortions in consumption levels across states ξ are minimized. See example in figure 3 when there is no rationing. There are two states $\varepsilon = H, \varepsilon = L$, with state-contingent prices p_H and p_L respectively. Price-insensitive consumers pay marginal price $p_L \leq p_\omega \leq p_H$. In each state ε, there is dead-weight loss (shaded in gray) as long as $p_\omega \neq p_\xi$, and marginal price p_ω is determined such that the expected dead-weight loss is minimal.

On top of that, noteworthy that in general, price-insensitive consumers are not faced with expected real-time prices, $p_\omega \neq p_\omega^F = \mathbb{E}_{\xi|\omega}(p_\xi)$. One exception would be no rationing happens, demand function $D_\varepsilon(p_\omega)$ is linear and demand shock is additive, so the first-order condition for price p_ω boils down to

$$\mathbb{E}_{\xi|\omega}(p_\omega - p_\xi) = 0 \Rightarrow p_\omega = p_\omega^F \tag{17}$$



(a) Linear inverse demand $p = \varepsilon - bD$

(b) Nonlinear inverse demand $p = p(D)$

Figure 3: Deadweight Loss of Price-insensitive Consumers

Note: This figure illustrates the determination of the optimal marginal price p_ω faced by price-insensitive consumers when there is no rationing. There are two states $\varepsilon = L, \varepsilon = H$, occurring with probability f_L and f_H, $f_L + f_H = 1$. State-contingent prices are p_L and p_H respectively. Price-insensitive consumers pay marginal price $p_L \leq p_\omega \leq p_H$. In each state ε, there is dead-weight loss (shaded in gray) as long as $p_\omega \neq p_\xi$, and optimal marginal price p_ω is determined such that the expected dead-weight loss is minimized. In panel (3a), the result is that $p_\omega = f_L p_L + f_H p_H$, which does not hold in a general case shown in panel (3b) when demand function is nonlinear.

Proposition 1 (b) suggests that price-sensitive consumers are faced with real-time price p_ξ and 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 (c) shows that for scheduled downward reserves, only technologies for which the marginal production cost larger than the real-time price are activated. And for scheduled upward reserves, only technologies for which the marginal production cost smaller than the real-time price are activated. Note that in real time, neither capacity investment nor flexibility commitment is reversible, so 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 ($T = 2$). Note that in state ω, marginal capacity employing technology θ have four exclusive commitment options: (1) idleness (2) energy, (3) energy plus downward reserve (4) upward reserve.[22]

For the sake of clarification, we define the following profit terms:

$$\pi^Q_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta \tag{18}$$

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

$$\pi^U_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[\max\{p_\xi - c^U_\theta, 0\}] - I^U_\theta \tag{20}$$

$\pi^Q_{\theta,\omega}$ is the marginal profit earned by technology θ if they commit to production at state ω. If this marginal capacity is also committed to provide downward flexibility, the net value of put option is referred as $\pi^D_{\theta,\omega}$. On the other hand, firms can choose to postpone production until the state ξ is observed, so the marginal capacity is provided as an upward reserve.

First, we discuss the decision on downward flexibility investment in the presence of production commitments, $Q_{\theta,\omega} \geq 0$. The put option value is

$$\pi^D_{\theta,\omega} = \mathbb{E}_{\xi|\omega}[\max\{c^D_\theta - p_\xi, 0\}] - I^D_\theta = \mu_{\theta,\omega} - \varphi^D_{\theta,\omega} \tag{21}$$

Hence, marginal downward reserve is valuable iff $\pi^D_{\theta,\omega} \geq 0$, and the shadow price of downside reserve is $\mu_{\theta,\omega} = \pi^D_{\theta,\omega}$, also termed *downside flexibility premium*. When $\pi^D_{\theta,\omega} < 0$, $\mu_{\theta,\omega} = 0$.

Then, we explain how the marginal capacity should be utilized between energy provision and upward flexibility. The profit of upward reserve is given by $\pi^U_{\theta,\omega}$, while to schedule energy gives profit $\pi^Q_{\theta,\omega} + \mu_{\omega,\omega}$. Hence, the marginal capacity is

[22] There is also a trivial case that the firm is indifferent among commitment options.

scheduled as upward reserve instead of energy iff

$$\pi^U_{\theta,\omega} \geq \pi^Q_{\theta,\omega} + \mu_{\theta,\omega} \tag{22}$$

The shadow price of capacity constraint $\lambda_{\theta,\omega}$ is determined such that the capacity is valued most. In other words, it maximizes the net value of commitment. Note that if neither energy nor reserve gives short-term positive profit, this marginal capacity should not be employed and the shadow price is zero. We summarize the commitment decision in table 2.

Profit Comparison	$\Delta Q_{\theta,\omega}$	$\Delta R^D_{\theta,\omega}$	$\Delta R^U_{\theta,\omega}$
$\pi^Q_{\theta,\omega} > \pi^U_{\theta,\omega}, \pi^D_{\theta,\omega} < 0, \pi^Q_{\theta,\omega} \geq 0$	+	0	0
$\pi^U_{\theta,\omega} < \pi^Q_{\theta,\omega} + \mu_{\omega,\omega} > 0, \pi^D_{\theta,\omega} \geq 0$	+	+	0
$\pi^U_{\theta,\omega} \geq \pi^Q_{\theta,\omega} + \mu_{\omega,\omega}, \pi^U_{\theta,\omega} \geq 0$	0	0	+
Otherwise	0	0	0

Table 2: Commitment Decision

We clarify the decision procedure with a simple example. Assume one firm has unit capacity, with marginal cost $c = 1, c^U = 5/4, c^D = 1/2$, and flexibility commitment cost $I^U = 1/2, I^D = 1/5$. The price tomorrow can be either $p_L = 0$ or $p_H = 4$. We consider four states of $\omega \in \{1, 2, 3, 4\}$.

(i) $\omega = 1, f_L = 1/4, f_H = 3/4$. That is, the probability of being in a low state is $1/4$ and a $3/4$ probability of being in a high state. If the firm chooses to sell it at forward price $p_F = 3$, the profit without downward flexibility is $\pi^Q = 2$. Additionally, the firm can invest in downward flexibility, which is utilized at low state, and recoups half of production cost. However, as low state probability is small, with investment cost I^D, the downside flexibility adds negative value $1/4 * (1/2 - 0) - 1/5 = -3/40$. Alternatively, the firm choosing upward reserve only produces when $p_H = 4$, and the expected profit is $\pi^U = 3/4 * (4 - 5/4) - 1/2 = 25/16$. As $25/16 < 2$, commitment to production and no flexibility provision would be the best choice. The shadow price of capacity constraint is $\lambda_1 = \pi^Q = 2$.

(ii) $\omega = 2, f_L = f_H = 1/2$. The expected spot price becomes $p_F = 2$, and the profit without downward flexibility is $\pi^Q = 1$. Because of a higher probability of low state, downside flexibility turns out to be profitable, $1/2 * (1/2 - 0) - 1/5 = 1/20$. The expected profit of an upward reserve becomes

$\pi^U = 1/2 * (4 - 5/4) - 1/2 = 7/8 < 1 + 1/20$. Hence, the firm should commit to production and at the same time, invest in downward flexibility. The shadow price of capacity constraint is $\lambda_2 = \pi^Q + \pi^D = 21/20$.

(iii) $\omega = 3, f_L = 3/4, f_H = 1/4$. As $p_F = 1$, the quantity commitment gives zero profit, but the accompanied downside flexibility further increases $3/4 * (1/2 - 0) - 1/5 = 7/40$. On the other hand, the upward reserve gives profit $\pi^U = 1/4 * (4 - 5/4) - 1/2 = 3/16 > 7/40$. Hence, the firm should schedule this capacity as upward reserve. The shadow price of capacity constraint is $\lambda_3 = \pi^U = 3/16$.

(iv) $\omega = 4, f_L = 9/10, f_H = 1/10$. Even upward flexibility is too expensive to be profitable, suggesting the firm should not use this unit capacity. The shadow price of capacity constraint is $\lambda_4 = 0$.

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

Now, we briefly explain how the equilibrium is determined numerically in a general scenario.[23] Given the parameters, we start with an arbitrary price distribution $F^0(p_\xi|\omega)$[24] for each ω. In the supply side, $\lambda_{\theta,\omega}$ as well as its expectation $E_\omega(\lambda_{\theta,\omega})$ is determined. The iteration continues until the free-entry condition is satisfied. If we set cost parameters arbitrarily, some investment opportunities can be dominated, $K_\theta = 0$. We rule out these investment opportunities as follows:

First, to list capacity investment cost I_θ in ascending order. Then do the same thing for expected shadow price $E_\omega(\lambda_{\theta,\omega})$. If the orders are exactly the same for all technologies, we admit all technologies and solve equilibrium prices from up to down. Otherwise, delete the technologies breaking the ordering and solve the equilibrium with remaining ones. For examples, if the ordering of investment cost of technology 1, 2, 3 is $I_1 < I_2 < I_3$ while the expected shadow price is $E(\lambda_3) < E(\lambda_1) < E(\lambda_2)$. Technology 3 must be strictly dominated and in equilibrium, only technology 1 and 2 are used. Solve the price when only technology 1 is used and then the price when both technology 1 and 2 are used.

Denote the equilibrium price distribution as $F^*(p_\xi|\omega)$. Equilibrium rationing rate α_ξ, marginal retail price p_ω, demand for price-sensitive consumers $\hat{\mathcal{D}}_\varepsilon$ demand, demand for price-insensitive consumers $\mathcal{D}_\varepsilon$ and commitment scheduling

[23] In section 5, we solve a closed form solution for a two-technology and binary state case.
[24] $F(p_\xi|\omega) = F_\omega(p_\xi)$. We use these two notations interchangeably.

in each state are solved automatically. Optimal investment distribution is then determined by the market clearance condition.

The trade-off between commitment alternatives depends on the commitment costs, adjustment costs, as well as state distribution. But for some special cases, the decisions are certain. See Lemma 1.

Lemma 1. *For special technologies,*

(i) Fully flexible technology is free to choose upward flexibility or production commitment accompanied with downward flexibility.

(ii) Fully inflexible technology commits to production if profitable.

(iii) If $I^D_\theta = 0$, energy provision is always accompanied with downward flexibility.

Lemma 1(i) is easy to verify by the first-order conditions presented in Proposition 1(d). When $c^U_\theta = c^D_\theta = c_\theta$ and $I^U(\theta) = I^D(\theta) = 0$. We directly derive:

$$\begin{aligned}&\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta, 0\}] > p^F_\omega - c_\theta\\ &\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta, 0\}] \geq 0 \quad \text{and equality holds if} \quad \forall \xi, p_\xi < c_\theta\end{aligned}$$

Lemma 1(ii) is straightforward that an inflexible firm decides to produce iff

$$\mathbb{E}_{\xi|\omega}[p_\xi] - c_\theta \geq 0$$

Lemma 1(iii) is claimed by the fact that $c^D_\theta \leq c_\theta$ and

$$\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c^D_\theta, 0\}] - [c_\theta - c^D_\theta] \geq p^F_\omega - c^D_\theta - [c_\theta - c^D_\theta] = p^F_\omega - c_\theta$$

Compared to a two-stage peak load pricing model, real-time prices in the presence of inflexibility exhibit two new features. Firstly, they can fall below the marginal generation cost. Moreover, they display higher volatility.

Lemma 2. *In state ω, the existence of downward reserve service implies the lower bound of distribution $F(p_\xi|\omega)$ is smaller than the production cost of the last employed technology; negative prices are possible when curtailment cost is larger than production cost, $c^D_\theta < 0$.*

$$R^D_\omega > 0 \Rightarrow \exists\{\xi, \theta\}, \quad p_\xi < c_\theta, q_{\theta,\xi} > 0$$

Proof. By Proposition 1(d) , if $\forall \xi, \theta, p_\xi > c_\theta$, $\mu^D_{\theta,\omega} = 0$. Hence, $R^D_{\theta,\omega} = 0 \Rightarrow R^D_\omega = 0$. □

Electricity prices fall with renewable penetration, which increases the value of downward flexibility and leads to a more frequent ramp-down of non-intermittent generation. Lack of decremental flexibility would result in below-cost and even negative prices. One recent example is the increasing frequency of negative prices in 2020 when supply from renewable output is higher but demand is low due to the Covid-19 pandemic.[25] Limitation of downward flexibility is witnessed in practice. For example, to cool down a nuclear plant, which usually serves as baseload, needs 10-20 hours and it takes more than half day to reach full load again. Also, the growing curtailment of solar energy in the afternoon demonstrates the challenge of ramping down thermal plants. Furthermore, to frequently change the operational rate needs a higher level of maintenance, which increases the operational and maintenance (O&M) cost. Finally, not only physical constraints but pricing policy curbs downside flexibility. For instance, wind power is easy to curtail but firms have no incentive to do it because subsidy is given to production from renewable sources.

Lemma 3. *Real-time price volatility decreases with flexibility:*

$$Var(p_\xi)[\textit{inflexibility}] \geq Var(p_\xi)[\textit{no inflexibility}] \tag{23}$$

This result is obvious by comparing different technologies in figure 2 as real-time supply elasticity increases with flexibility. Hence, system flexibility helps to reduce market volatility. We elaborate more insights into flexibility in the following part.

3.5 Flexibility premium

As shown in Lemma 1, flexible and inflexible technologies have different strategies available. Inflexible technologies make production commitment if the difference between expected price p^F_ω and their own production cost c_θ is positive, $\pi_{\theta,\omega} = \pi^Q_{\theta,\omega} \geq 0$. Flexible ones have a wider range of choices. It is easy to check that a firm with fully flexible technology always chooses to take advantage of its flexibility and earns a higher profit than that with an inflexible one, given they have the same marginal production cost.

Lemma 4. *For any pair of technology* (θ_1, θ_2) *such that* $c_1 = c_2$*, and* $I^U_1 = I^D_1 =$

[25] For instance, Ireland and Germany having large share of wind generation saw negative day-ahead prices 4.2% and 3.4% respectively of 2020.

$0, c_1^U = c_1^D = c_1$,

$$\forall \omega, \quad \pi_{1,\omega} \geq \pi_{2,\omega}$$

More generally, we obtain

Proposition 2. *In state ω,*

- *active inflexible firms earn expected real-time price $p_\omega^F, \pi_{\theta,\omega} = p_\omega^F - c_\theta$.*
- *active flexible firms earn a technology-specific premium $\nu_{\theta,\omega} \geq 0$* [26] *on top of the opportunity cost of not selling at expected real-time price*

$$\pi_{\theta,\omega} = p_\omega^F - c_\theta + \nu_{\theta,\omega} \tag{24}$$

Proof.

$$\pi_{\theta,\omega} = \max\{\pi_{\theta,\omega}^Q, \pi_{\theta,\omega}^D + \pi_{\theta,\omega}^Q, \pi_{\theta,\omega}^U, 0\} \geq \pi_{\theta,\omega}^Q = p_\omega^F - c_\theta$$

□

The prevailing consensus suggests that upward reserve should be made up for the opportunity cost of not serving energy in the day-ahead market. We argue that this is not true. If upward flexibility is scheduled, the marginal profit is $\pi_{\theta,\omega}^U$, and $\pi_{\theta,\omega}^Q$ is referred to as the *opportunity cost*. Since

$$\pi_{\theta,\omega}^U + \varphi_{\theta,\omega}^U = \pi_{\theta,\omega}^Q + \mu_{\theta,\omega} + \varphi_{\theta,\omega}^Q \tag{25}$$

$\Delta\pi_{\theta,\omega}^U = \pi_{\theta,\omega}^U - \pi_{\theta,\omega}^Q = \mu_{\theta,\omega} + \varphi_{\theta,\omega}^Q - \varphi_{\theta,\omega}^U$ is the difference of profits between energy provision and upward reserve. When the marginal unit is used as upward reserve, $\Delta\pi_{\theta,\omega}^U = \nu_{\theta,\omega} = \mu_{\theta,\omega} + \varphi_{\theta,\omega}^Q \geq 0$ is *upward flexibility premium*.

This premium consists of two parts. First, bidding in the day-ahead energy market does not exclude the possibility of providing downward flexibility. Hence, providing upward reserve not only incurs the opportunity cost of not serving energy but also the accompanied profit from the downward reserve $\mu_{\theta,\omega}$. Then, the non-negative constraint implies $\varphi_{\theta,\omega}^Q > 0$, the shadow price of take long energy position. Note that $R_{\theta,\omega}^D \leq R_{\theta,\omega}^Q$ and $R_{\theta,\omega}^D \geq 0$ imply $R_{\theta,\omega}^Q \geq 0$, so $\varphi_{\theta,\omega}^Q$ and $\mu_{\theta,\omega}$ are not determined separately. The sum is fixed since there is only one degree of freedom.

This result serves as the starting point for the subsequent analysis of the impacts of market design on flexibility investment. The key insight of real options

[26] Decomposition of premium is shown in Appendix B.

theory is that waiting can be valuable under uncertainty when the value of information outweighs the cost of postponement. In this paper, waiting creates production flexibility, by reducing the loss when prices are low. As neither commitment nor adjustment is free, firms trade off the cost and value of flexibility. The remaining part in this section demonstrates that a firm with higher production cost, lower commitment cost, or lower adjustment cost is more likely to commit to flexibility, holding other things constant. Moreover, flexibility is more valuable when the future is less predictable. Flexibility premium of technology θ has an upper bound[27]:

$$\bar{\nu}_{\omega}(\theta) = c_{\theta}F(c_{\theta}|\omega) - \int_{\underline{p}}^{c_{\theta}} p_{\xi}dF(p_{\xi}|\omega) < c_{\theta} - \underline{p}$$

It is clear that the maximal flexibility premium increases with production cost and decreases with the lower bound of market prices. I emphasize this observation in the following lemmas.

Lemma 5. $\forall c_{\theta}, \omega$

- *Ceteris paribus, flexibility value increases with production cost, and decreases with adjustment and flexibility investment costs:*

$$\frac{\partial \nu^{U}_{\theta,\omega}}{\partial c} = F_{\omega}(c^{U}_{\theta}) > 0; \frac{\partial \nu^{U}_{\omega}}{\partial (c^{U}_{\theta} - c_{\theta})} = -[1 - F_{\omega}(c^{U}_{\theta})] < 0; \frac{\partial \nu^{U}_{\theta,\omega}}{\partial I^{U}_{\theta}} = -1 < 0$$
$$\frac{\partial \nu^{D}_{\theta,\omega}}{\partial c_{\theta}} = F_{\omega}(c^{D}_{\theta}) > 0; \frac{\partial \nu^{D}_{\theta,\omega}}{\partial (c_{\theta} - c^{D}_{\theta})} = -F_{\omega}(c^{D}_{\theta}) < 0; \frac{\partial \nu^{D}_{\theta,\omega}}{\partial I^{D}_{\theta}} = -1 < 0$$

- *A technology with higher production cost c can earn higher premium $\nu^{U}_{\omega}(\theta)$, by postponing production if and only if the weighted average increase of inflexibility cost is lower than the probability that reserves are not activated:*

$$\frac{dI^{U}_{\theta}}{dc_{\theta}} + \frac{dc^{U}_{\theta}}{dc_{\theta}}[1 - F_{\omega}(c^{U}_{\theta})] < 1 \tag{26}$$

  and inequality holds for sure if production cost c_{θ} is independent of flexibility investment cost I^{U}_{θ} and adjustment cost $c^{U}_{\theta} - c_{\theta}$.

- *A technology with higher production cost c_{θ} can earn higher premium $\nu^{D}_{\theta,\omega}$, by investing in downside flexibility if and only if the weighted average increase*

[27] This upper bound is derived by assuming $I^{U}_{\theta} = I^{D}_{\theta} = 0, c^{D}_{\theta} = c^{U}_{\theta} = c_{\theta}$. That is, when technology θ is fully flexible.

of inflexibility cost is lower than the probability that reserves are activated:

$$\frac{dI_\theta^D}{dc_\theta} - \frac{dc_\theta^D}{dc_\theta} F_\omega(c_\theta^D) \leq 0 \tag{27}$$

and inequality holds if production cost c_θ is independent of flexibility investment cost I_θ^D and adjustment cost $c_\theta - c_\theta^D$.

Flexibility premium monotonically decreases with inflexibility cost, ceteris paribus. However, since higher production cost increases the value of flexibility, when all of the operational costs are positively correlated, the net effect is determined by (26) if offering upward reserves and (27) for downward reserves. This establishes a counter-intuitive statement that a more costly technology is possible to earn a higher premium: a less flexible technology can earn flexibility value more than a flexible one, when production cost of this less flexible firm is higher, because the technology can benefit more from not producing when price cannot cover cost.

However, it is worth bearing in mind that a higher premium does not imply a higher profit. A firm with higher production and inflexibility costs is possible to earn a higher premium, but the profit must be lower.

Lemma 6. *Flexibility premium is a nonincreasing function of price distribution that exhibits second-order stochastic dominance. That is, if price distribution $F_\omega^1(p)$ second-order stochastically dominates $F_\omega^2(p)$,*

$$\nu_{\theta,\omega}^1 \leq \nu_{\theta,\omega}^2 \tag{28}$$

strict inequality holds when the real-time marginal cost is larger than the lowest price.

Proof. Note that only the composition $\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta^U, 0\}]$ of $\lambda_{\theta,\omega}^U$ and $\mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta^D, 0\}]$ of $\lambda_{\theta,\omega}^D$ depends on $F_\omega(p)$, and they share the same format $\mathbb{E}[\max\{x - x^*, 0\}] = E_{\xi|\omega}(p) - x^* + x^* F(x^*) - \int_{\underline{x}}^{x^*} x dF(x)$. Hence, the premium $\nu \propto \int_{\underline{x}}^{x^*} F(x)dx$[28]. By definition, distribution $F_\omega^1(x)$ has second-order stochastic dominance over $F_\omega^2(x)$ if and only if $\int_{\underline{x}}^{x^*} F_\omega^1(x)dx \leq \int_{\underline{x}}^{x^*} F_\omega^2(x)dx$. □

To put it differently, flexibility becomes more valuable when real-time prices are less predictable. Integration of intermittent renewables increases spot price volatility and requests more system flexibility. Lemma 6 implies an increase of

[28] x^* is a constant. $x^* = c_\theta^U$ for $\lambda_\omega^U(\theta)$, and $x^* = c_\theta^D$ for $\lambda_\omega^D(\theta)$.

premium of flexible assets that operate mostly for providing flexibility alongside energy transition. A price signal that correctly reflects this premium is vital to flexibility operation as well as investment.

Note that a second-order stochastically dominated distribution $F^2_\omega(x)$ has mean not greater than $F^1_\omega(x)$. Hence, it is possible that with higher renewable penetration, some expensive technologies become idle so they neither commit to production nor flexibility. However, as we stick to the definition that flexibility premium is the difference of profit between a flexible and inflexible technology with the same production cost, Lemma 6 still holds.

Finally, as first-order dominance is a sufficient condition for second-order dominance, Lemma 6 implies that flexibility becomes more valuable when prices are more likely to be low, which is the case when more renewables are integrated into the power system. As stated before, the advantage of flexibility over production commitment is to provide flexibility of *not* producing when prices are low.

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 market design on flexibility investment and social welfare.

4.1 Real-time Only Market

According to the first fundamental theorem of welfare economics, the competitive equilibrium where market clears is equivalent to social optimum, which also applies in this paper. In a decentralized real-time market, generators are paid real-time price p_ξ. Retailers and price-sensitive consumers face p_ξ and price-insensitive consumers are charged a two-part tariff with fixed fee A and marginal price p_ω. Also, retailer are allowed and able to ration price-insensitive consumers by α_ξ. The profit maximization problem of a generator with technology θ is

$$\begin{aligned} \max_{Q_{\theta,\omega}, R^U_{\theta,\omega}, R^D_{\theta,\omega}} \quad & \mathbb{E}[p_\xi q_{\theta,\xi} - C_{\theta,\xi}] \\ \text{s.t. } \forall \xi, \quad & Q_{\theta,\omega} + R^U_{\theta,\omega} \le K_\theta \\ & R^D_{\theta,\omega} \le Q_{\theta,\omega} \end{aligned} \tag{29}$$

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

p_ω and fraction of satisfied demand α_ξ to maximize the consumers' surplus (recall that retail market is competitive and retailer's profit is zero)

$$\mathbb{E}[\mathcal{S}_\varepsilon(p_\omega, \alpha_\xi) - p_\xi \mathcal{D}_\varepsilon(p_\omega, \alpha_\xi)] \tag{30}$$

which gives the conditions characterized by (a). Therefore, real-time market can price in flexibility and maximize social efficiency.

4.2 Day-ahead Only Market

Notwithstanding the sufficiency of real-time market to convey correct price signals for flexibility operation and investment, it is hard to coordinate large amount of trade just before delivery. One solution is to build a day-ahead forward market to trade energy. The no-arbitrage condition requires that the equilibrium day-ahead price for each state is equal to the expected real-time prices such that energy supplier is indifferent between trading in the day-ahead market and real-time market, while flexibility providers weakly prefer to trade in the real-time market. In practice, demand information is revealed overtime, so the supply strategies of more flexible firms might change and they are more willing to give up their flexibility and sell energy. Hence, it is rational to adjust energy purchase continuously through the intra-day market, and only the most flexible firms wait until real-time market opens.

Therefore, even a real-time market functions well in theory, day-ahead and intra-day markets are justified to coordinate the supply and reduce the burden of real-time computation and trading. To the contrary, when flexible firms have to trade before demand is realized and they are not able to adjust in a real-time market, they lose the advantage of flexibility, inducing both production and investment distortion.

Proposition 3. *In the absence of real-time market, a day-ahead energy only market would result in under-investment in flexible technologies and over-investment in inflexible technologies; the equilibrium day-ahead price for each state is characterized by*

$$\mathbb{E}_\omega(p^F_\omega) = \mathbb{E}_\omega[\mathbb{E}_{\xi|\omega}(p_\xi)] \tag{31}$$

and $p^F_\omega = \mathbb{E}_{\xi|\omega}(p_\xi)$ *in a continuum setup.*

Proof. When there is only a day-ahead market, the maximization problem be-

comes

$$\begin{aligned} &\mathbb{E}\{S_\varepsilon(Q_\omega) - C^Q_\omega\} - C^I \\ &\text{s.t. } \forall\omega, \quad Q_{\theta,\omega} \leq K_\theta \end{aligned} \tag{32}$$

The results give

$$\mathbb{E}_\omega \max\{\mathbb{E}_{\xi|\omega}[S'_\varepsilon(Q_\omega)] - c_\theta, 0\} = I_\theta \tag{33}$$

$$p^F_\omega = \mathbb{E}_{\xi|\omega}[S'_\varepsilon(Q_\omega)] \tag{34}$$

Eq.(33) is the long-run equilibrium condition for technologies that do not provide any flexibility[29]. Recall that when there is a real-time market, an inflexible technology has free-entry condition:

$$\mathbb{E}_\omega \max\{\mathbb{E}_{\xi|\omega}(p_\xi) - c_\theta, 0\} = I_\theta \tag{35}$$

Since there is one technology serves base load, $\mathbb{E}_\omega(p^F_\omega) = \mathbb{E}_\omega[\mathbb{E}_{\xi|\omega}(p_\xi)]$. If $p^F_\omega = \mathbb{E}_{\xi|\omega}(p_\xi)\forall\omega$, no flexible technologies exist in the long run. In other words, existence of flexibility implies $p^F_\omega > \mathbb{E}_{\xi|\omega}(p_\xi), \exists\omega$.

(i) When there is a continuum of states and technologies or the number of feasible technologies is larger than the number of states for ω. The distribution of p^F_w and $\mathbb{E}_{\xi|\omega}(p_\xi)$ must be the same to satisfy (33) and (35) simultaneously. That is, $p^F_w = \mathbb{E}_{\xi|\omega}(p_\xi)$ should hold everywhere, and no flexible capacity is invested.

There is at least one state ω such that all inflexible capacity K^I is used: $p^F_\omega = \mathbb{E}_{\xi|\omega}[S'_\varepsilon(K^I)]$. As $S''_\varepsilon \leq 0$, the total inflexible capacity must increase compared to optimum.

(ii) On the other hand, if the number of technologies available is less than the number of states, and in equilibrium, $p^F_\omega > \mathbb{E}_{\xi|\omega}(p_\xi), \exists\omega$, as $\mathbb{E}_\omega(p^F_\omega) = \mathbb{E}_\omega[\mathbb{E}_{\xi|\omega}(p_\xi)]$ still holds, $p^F_\omega = \mathbb{E}_{\xi|\omega}[S'_\varepsilon(K^I)] \leq \mathbb{E}_{\xi|\omega}(p_\xi), \exists\omega$, and the total inflexible capacity must increase compared to optimum.

There exists a state ω when all technologies are used and $p^F_\omega > \mathbb{E}_{\xi|\omega}(p_\xi)$, so total demand and flexible supply should decrease compared to optimum.

□

[29] A technology does not provide flexibility because it is technically inflexible or it is always beneficial to be scheduled in advance. For simplification, we call both "inflexible" technology.

The result implies that forward bias can come from the lack of efficient real-time market,[30] when the number of technologies is much less than the number of states $|\Omega|$. However, in our paper, we focus on the case $p^F_\omega = \mathbb{E}_{\xi|\omega}(p_\xi)$.

The arguments between day-ahead and real-time energy markets are often debated. Most literature focuses on day-ahead market, regarding it as a way to hedge price volatility(Bessembinder and Lemmon, 2002) and to mitigate market power by allowing suppliers to trade in forward markets and chase a leading position (Allaz and Vila, 1993; Puera and Bunn, 2022). This paper is developed within the framework of perfect competition and risk neutrality. It does not mean the real-time market is always superior to the day-ahead market, but appeals more attention to the instantaneous one as it internalizes the cost of flexibility and conveys correct price signal that fails to be properly reflected in the day-ahead market.

4.3 Reserve Market

This part shows how a reserve market complements the day-ahead energy market and restores efficiency as in a real-time market. In this paper, there is no needs for risk hedging. Hence, a reserve market is redundant if a well-organized real-time market exists. However, when real-time settlement is not possible, Appendix D shows that the reserve market can behave as an options market to provide flexibility and therefore to obtain efficient investment in flexibility.

Proposition 4. *In the absence of a real-time market, the social optimum can be replicated by a day-ahead market with uniform price $p^F_\omega = E_{\xi|\omega}(p_\xi)$, and in each state ω, a reserve market providing technology-specific menu $M = \{p^{K,U}_\theta, p^{K,D}_\theta, p^{X,D}_\theta, p^X_D(\theta)\}$, where:*

$$
\begin{aligned}
p^{K,U}_\theta &= p^F_\omega - c_\theta + I^U_\theta + \nu_{\theta,\omega}, \textit{ is upward capacity payment}\\
p^{K,D}_\theta &= I^D_\theta + \nu_{\theta,\omega}, \textit{ is downward capacity payment}\\
p^{X,U}_\theta &= c^U_\theta, \textit{ is activation payment for upward reserves}\\
p^{X,D}_\theta &= c^D_\theta, \textit{ is activation payment for downward reserves}
\end{aligned}
$$

Hence, both a real-time market and a day-ahead forward market combined with a reserve market are equivalent. A real-time market is a decentralized market and excels in single pricing, since a technology-specific payment is hard to implement

[30] The literature shows premium is a result of risk hedging (Bessembinder and Lemmon, 2002) or market power (Ito and Reguant, 2016).

in practice and it requires centralized procurement, while it is also challenging in quick response to demand realization.

The result in proposition 4 is consistent with Chao and Wilson (1987) in which they show that in a retail market, 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. In this paper, generators know their own cost structure and select the contracts for providing flexibility service. 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 price can be replaced by a day-ahead forward price for inflexible technologies, combined with a menu of reserve contracts for flexible ones.

It should be noted that production is not always flexible, forward is not a special option with zero activation payment since negative prices are possible. One may argue that we can set activation price as negative as possible and adjust capacity payment accordingly for such an option. This is theoretically correct, but not a common way to design options. Also, it means in some cases, call and put options have to be activated at the same time. More importantly, inflexible firms are not able to adjust so it is not reasonable that they sell an option. Therefore, we state that social efficiency is obtained through a real-time market, or a combination of a day-ahead forward market and a reserve market.

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 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 day-ahead price and its 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 day-ahead clearing price is F_ω^F. A firm with production cost c_θ has expected profit[31]:

$$\pi_{\theta,\omega} = \int_{b_\theta}^{\infty} (p_\omega^F - b_\theta) dF_\omega^F + (b_\theta - c_\theta)(1 - F_\omega(b_\theta))(1 - F_\omega^F(b_\theta)) \tag{36}$$

The optimal bidding is given by:

$$b_\theta^o = c_\theta - \frac{[1 - F_\omega^F(b_\theta^o)]F_\omega(b_\theta^o)}{F_\omega(b_\theta^o)[1 - F_\omega^F(b_\theta^o)] + F_\omega^F(b_\theta^o)[1 - F_\omega(b_\theta^o)]} \tag{37}$$

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

$$\pi_{\theta,\omega}^e = \int_{c_\theta}^{\infty} (p_\xi - c_\theta) dF_\omega \tag{38}$$

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

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

$$\pi_{\theta,\omega} = (p_\omega^F - b_\theta) + (b_\theta - c_\theta)(1 - F_\omega(b_\theta)) \tag{39}$$

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

$$b_\theta^e = c_\theta - \frac{\int_{-\infty}^{c_\theta} (c_\theta - p_\xi) dF_\omega}{F_\omega(b_\theta^e)} \neq b_\theta^o \tag{40}$$

□

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

[31] 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.

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 bid. Chao and Wilson (2002) proposes 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 real-time price and capacity bid should be the difference of foregone profit in day-ahead energy price and the expected profit from called energy paid the real-time price. Therefore, it is not surprising that the optimal bid is negative in the example of that paper. We prove that negative bid is not a special case, but an inevitable result of 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 to have a capacity market for reserves.

When capacity and activation price are both predetermined, uniform pricing for reserves cannot achieve long-term optimum even 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

$$p^{K,U}_\omega - I^U_\theta + [1 - F_\omega(c^U_\theta)][p^{X,U}_\omega - c^U_\theta] = \pi_{\theta,\omega}, \quad \forall\theta \tag{41}$$

for upward reserves, and

$$p^{K,D}_\omega - I^D_\theta + F_\omega(c^D_\theta)[p^{X,D}_\omega + c^D_\theta] = \nu_{\theta,\omega}, \quad \forall\theta \tag{42}$$

for downward reserves. Hence, the system of linear equations in (41) or (42) has a unique set of solution 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 day-ahead energy market and direct costs to provide reserves if any as flexibility premium behaves as a reward instead of 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 flexibility premium if the system operator does not pay an extra payment, and a recursive inference shows no flexible assets exist in the long run.

Lemma 8. *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 day-ahead 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 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

$$\mathbb{S}_\omega = b^K + \int_0^{b^X} [1 - F_\omega(p_\xi)] dp_\xi \tag{43}$$

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

$$\mathbb{E}(\pi(b_\theta^K, b_\theta^X)|\theta) = [b_\theta^K - (p^F - c_\theta + I_\theta^U)]\mathbb{P}_\omega(\mathbb{S}_\omega(b_\theta^K, b_\theta^X), \mathbb{S}_{-i}) \tag{44}$$

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

$$b_{\theta,\omega}^{K,o} = -\frac{1}{\ln \mathbb{P}_\omega[\mathbb{S}_\omega(b_{\theta,\omega}^{K,o}, c_\theta^U)]'} + p^F - c_\theta + I_\theta^U \tag{45}$$

where

$$\ln \mathbb{P}_\omega[\mathbb{S}_\omega(b^{K,o}_{\theta,\omega}, c_\theta)]' = -\frac{\partial \mathbb{P}_\omega[\mathbb{S}_\omega(b^{K,o}_{\theta,\omega}, c_\theta)]/\partial \mathbb{S}_\omega(b^{K,o}_{\theta,\omega}, c_\theta)}{\mathbb{P}_\omega[\mathbb{S}_\omega(b^{K,o}_{\theta,\omega}, c_\theta)]} \tag{46}$$

A direct result is $b^{K,o}_{\theta,\omega} \geq p^F - c_\theta + I^U_\theta$, 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 to win 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 short-term capacity cost. To equalize information rent and flexibility premium, the probability of being selected $\mathbb{P}_\omega$ is given by

$$\mathbb{P}_\omega(\mathbb{S}_\omega) = e^{-\int_0^{\mathbb{S}_\omega} \frac{1}{V_\omega(t)} d_t} \tag{47}$$

where $\nu_\omega = V_\omega(\mathbb{S})$. One necessary assumption is that flexibility premium is a function of score, so we rule out the possibility that flexible firms with different flexibility value have the same score in equilibrium. However, by Lemma 5, 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 flexibility premium, the effects are perfectly cancelled out. Hence, for each state ω, at most one type of reserve provider is able to earn efficient rent.

Lemma 9. *In a pay-as-bid scoring auction with scoring rule $\mathbb{S}_\omega$ is given by Eq.(43), 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 day-ahead energy market plus information rent as capacity bid, which is given by Eq.(45). 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 menu can be a good compromise if real-time settlement is not possible. However, this market is demanding in designing as they depend on the time-varying data specified by the system operator and hence highly sensitive to the information collected and 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 premium. It is pivotal that each generator chooses among different level of capacity payments in exchange for being available in real-time at 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:*

$$W(RT) = W(menu) \geq W(scoring) = W(Integrated) = W(uniform) = W(DA\ only) \tag{48}$$

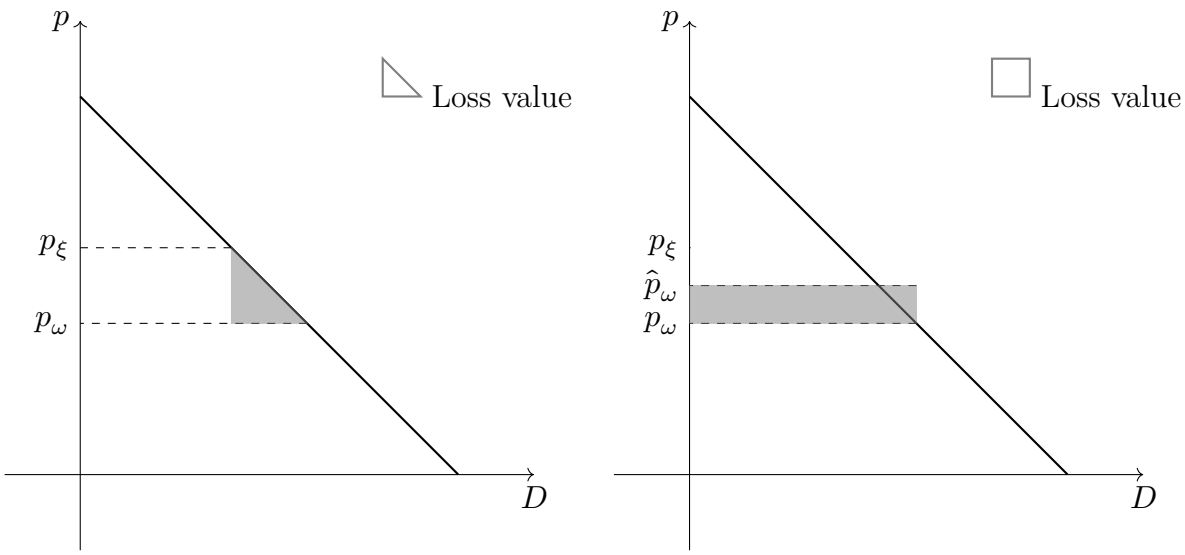
4.5 Demand Flexibility

Rationing The results in ((a)) and ((b)) 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 10. *(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: $VOLL_\xi = \frac{S_\varepsilon(p_\omega)}{D_\varepsilon(p_\omega)} = \hat{p}_\omega \geq p_\omega$; rationing happens when real time price is larger than $VOLL_x i$: $p_\xi > \hat{p}_\omega \Leftrightarrow \alpha_\xi = 0$.*

Share of inflexible consumers The first-order conditions in Proposition 1(a) 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.*



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

Figure 4: Rationing for Price-insensitive Consumers

Proposition 6 is immediately proven by lemma 6 and 10. 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 ($\hat{p}_\omega < p_\xi$), 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 ($p_\omega < p_\xi < \hat{p}_\omega$). 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 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 Example

This section gives an example to quickly understand the key points of this paper. 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 $I_2^U = I_2^D = c_2^U = |c_2^D| = \infty$[32]. The investment cost of technology 1 is larger than that of technology 2: $I_1 > I_2$. All consumers can react to real-time prices ($\theta = 0$). The total capacity of each technology is denoted by K_1, K_2.

Denote real-time prices for both states as p_L and p_H, respectively. The social planner's objective[33] 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 $q_{1,H}$:

$$\begin{aligned} \max_{\{K_1, K_2, q_{1,\varepsilon}, Q_2\}} \quad & \mathbb{E}[S_\varepsilon(q_{1,\varepsilon} + Q_2) - c \cdot (q_{1,\varepsilon} + Q_2)] - I_1 K_1 - I_2 K_2 \\ s.t. \quad & q_{1,\varepsilon} \leq K_1 \\ & q_{2,\varepsilon} = Q_2 \\ & Q_2 \leq K_2 \end{aligned} \tag{49}$$

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 $E(p) = c + I_2$;

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

[33] 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, and both types of firms need to determine the quantity and prices before demand is realized. The maximization problem becomes

$$\begin{aligned} \max_{\{K_1,K_2,Q_1,Q_2\}} \quad & E[S_\varepsilon(Q_1+Q_2) - c\cdot(Q_1+Q_2)] - I_1K_1 - I_2K_2 \\ s.t. \quad & Q_1 \le K_1 \\ & Q_2 \le K_2 \end{aligned} \tag{50}$$

which gives

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

p_F is the unit payment to generators, and p_R is the price charged to consumers. As predicted, in the absence of real-time markets and presence of a forward market, long-term equilibrium shows:

a) no investment in flexible technology;

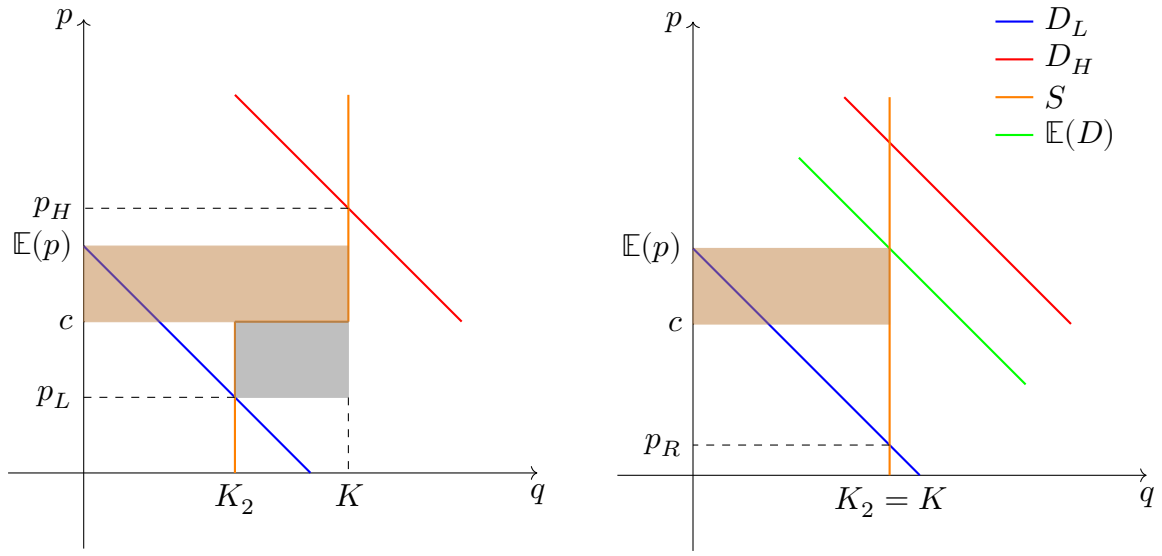
b) over-investment in inflexible technology;

c) rationing in high demand scenario;

d) fixed fee charged to consumers which amounts to $(p_F - p_R)K$.

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 firm and charges a two-part tariff from consumers.

Next, we show how an options market complements forward market and achieves 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 for unit quantity they sell in the forward market; if they trade in the options market, they



(a) Real-Time Market

(b) Day-ahead Only Market

Figure 5: Equilibrium in Real-time and Forward Market

Note: This figure represents the equilibrium comparison between a real-time market and a day-ahead forward only market. Sub-figure (5a) shows the equilibrium prices, quantities and investments when a well-functioning real-time market exists. In this case, firms making a production commitment earn expected price $\mathbb{E}(p)$ and the net profit is shade in brown. In addition, flexible firms can react to real-time demand shock so they avoid the loss at low state(shaded in gray). Flexibility premium is unconditional loss avoidance, so brown area times the probability of being in a low state p_L, which is equal to $I_1 - I_2$. Sub-figure (5b) shows the equilibrium when 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 low state and rationing in high state. $K_2(b) > K_2(a)$, and market clearance requires that $p_R < p_L < c$. Retailers are break-even through a fixed fee.

receive a capacity payment p_K for being available in real-time and a production payment p_X if the option being 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 forward price p_F is higher than production cost c. The objective function of flexible firms is

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

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:

$$p_K = f_H(p_H - p_X), \quad p_L \leq p_X \leq p_H \tag{51}$$

As shown in lemma 8, when there is only one technology providing flexibility, there are infinitely many combinations of strike and activation prices. If strike price is equal to production cost, $p_X = c$, a flexible firm should be paid a capacity payment p_K larger than opportunity cost of not trading in the forward market:

$$p_K > \mathbb{E}_\varepsilon(p) - c \tag{52}$$

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

6 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 short-run profit is negative. Second, in the absence of a real-time market, a forward only market would cause under-investment in flexibility and over-investment in inflexible technologies. Finally, an options market may help to restore the efficiency by an array of technology-specific contracts that compensate for flexibility.

This model is relevant to electricity markets. It indicates the importance of a real-time or balancing market to provide correct incentive 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 capacity market, arguing that flexibility providers need to reimburse their startup and adjustment cost via capacity market, which is not supported by this paper. We show that balancing price is able to reimburse any direct and indirect costs associated with adjustment. Therefore, when there is no risk-aversion and other market failure[34], a balancing capacity market is not

[34]In practice, risk aversion, market power and price cap can rationalize the capacity market,

necessary to complement balancing energy market.

Even this paper analyzes in the context of electricity market, flexibility premium 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 size of airplanes and hiring more people as reserves.

Last, 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 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 separate purchase of incremental and decremental reserves. Furthermore, the investment cost in flexibility is also reasonable that 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.

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

but reimbursement for direct adjustment payment is not a good reason.

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.

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

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

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

B Proof of Proposition ??

The shadow price $\lambda_\omega^U(\theta)$ of capacity constraint provided by technology θ is given by

$$\begin{aligned}
\lambda_\omega^U(\theta) =& \mathbb{E}_{\xi|\omega}[\max\{p_\xi - c_\theta^U, 0\}] - I^U(\theta) \\
=& \mathbb{E}_{\xi|\omega}[p_\xi] - \int_{\underline{p}}^{c_\theta^U} p_\xi dF_\omega - c_\theta^U[1 - F_\omega(c_\theta^U)] - I_\theta^U \\
=& \lambda_{\theta,\omega}^Q + \underbrace{c_\theta F_\omega(c_\theta) - \int_{\underline{p}}^{c_\theta} p_\xi dF_\omega}_{I} - \underbrace{\{\int_{c_\theta}^{c_\theta^U} p_\xi dF_\omega - c_\theta[F_\omega(c_\theta^U) - F_\omega(c_\theta)]\}}_{II} \\
& - \underbrace{[c_\theta^U - c_\theta][1 - F_\omega(c_\theta^U)]}_{III} - \underbrace{I_\theta^U}_{IV} \\
\nu_\omega =& \max\{I - (II + III + IV), 0\}
\end{aligned}$$

(i) I: benefit from not producing when price is lower than production cost;

(ii) II: regret cost because of short-term inflexibility;

(iii) III: incurred adjustment cost;

(iv) IV: investment cost in flexibility.

Similarly, $\lambda_\omega^D(\theta)$ is

$$\begin{aligned}
\lambda_\omega^D(\theta) =& \mathbb{E}_{\xi|\omega}[\max\{p_\varepsilon - c_\theta^D, 0\}] - [c - c_\theta^D] - I_\theta^D \\
=& \lambda_{\theta,\omega}^Q + \underbrace{c_\theta F_\omega(c_\theta^D) - \int_{\underline{p}}^{c_\theta^D} p_\xi dF_\omega}_{V} - \underbrace{[c_\theta - c_\theta^D]F_\omega(c_\theta^D)}_{VI} - \underbrace{I_\theta^D}_{VII} \\
\nu_\omega =& \max\{V - (VI + VII), 0\}
\end{aligned}$$

(i) V: benefit from not producing when price is low;

(ii) VI: incurred adjustment cost;

(iii) VII: investment cost in flexibility.

C Proof of Lemma 5

Proof. Define the difference of net values of upward reserves and energy by:

$$\Delta\lambda_{\theta,\omega} = \lambda^U_\omega(\theta) - \lambda^Q_\omega(\theta) \tag{C.1}$$

$$\Delta\lambda^U_\omega(\theta) = \int_{c^U_\theta}^{\bar{p}} p_\xi dF_\omega - c^U_\theta[1 - F_\omega(c^U_\theta)] - I^U_\theta - p^F_\omega + c_\theta \tag{C.2}$$

$$= -\int_{\underline{p}}^{c^U_\theta} p_\xi dF_\omega - c^U_\theta[1 - F_\omega(c^U_\theta)] - I^U_\theta + c_\theta \tag{C.3}$$

$$\Rightarrow \frac{d\lambda^U_\omega(\theta)}{dc_\theta} = 1 - \frac{dI^U_\theta}{dc_\theta} - \frac{dc^U_\theta}{dc_\theta}[1 - P(c^U_\theta)] \geq 0 \tag{C.4}$$

$$\Leftrightarrow \frac{dI^U_\theta}{dc_\theta} + \frac{dc^U_\theta}{dc_\theta}[1 - F_\omega(c^U_\theta)] \leq 1 \tag{C.5}$$

Similarly,

$$\Delta\lambda^D(\theta) = cP_\omega(c^D) - \int_{\underline{p}}^{c^D} p_\xi dF_\omega - [c - c^D]P_\omega(c^D) - I^D_\theta \tag{C.6}$$

$$\Rightarrow \frac{d\Delta\lambda^D}{dc} = -\frac{dI^D}{dc} + \frac{dc^D}{dc}F_\omega(c^D) \geq 0 \tag{C.7}$$

□

D Proof of Proposition 4

Proof. For any θ choosing upward reserves, incentive compatibility requires:

$$\begin{aligned}
\lambda^U(\theta) &= (p^K_{U\,\theta} - I^U_\theta) + [1 - F_\omega(p^X_U(\theta))](p^X_U(\theta) - c^U_\theta) \\
&\geq (p^K_U(\theta') - I^U_\theta) + [1 - F_\omega(p^X_U(\theta'))](p^X_U(\theta') - c^U_\theta) \\
&= \lambda^U(\theta') + I^U(\theta') - I^U_\theta + [1 - F_\omega(p^X_U(\theta'))][(p^X_U(\theta') - p^X_U(\theta)]
\end{aligned}$$

We propose $p^K_U(\theta) = p^F_\omega - c_\theta + I^U(\theta) + \nu_{\theta,\omega}, p^X_U = c^U_\theta$.

$$\begin{aligned}
&p^F_\omega - c_\theta + I^U(\theta) + \nu_{\theta,\omega} - I^U_\theta \\
\geq &p^F_\omega - c(\theta') + I^U(\theta') + \nu_\omega(\theta') - I^U_\theta + [1 - F_\omega(c^U(\theta'))][(c^U(\theta') - c^U_\theta]
\end{aligned}$$

As $\nu_\omega(\theta) = -\int_{\underline{p}}^{c^U_\theta} p_\xi dF_\omega - c^U_\theta[1 - F_\omega(c^U_\theta)] - I^U_\theta + c_\theta$. The inequality becomes:

$$
\begin{aligned}
&- \int_{\underline{p}}^{c_\theta^U} p_\xi dF_\omega - c_\theta^U[1 - F_\omega(c_\theta^U)] \\
&\geq - \int_{\underline{p}}^{c^U(\theta')} p_\xi dF_\omega - c^U(\theta')[1 - F_\omega(c^U(\theta'))] + [1 - F_\omega(c^U(\theta'))][(c^U(\theta') - c_\theta^U]
\end{aligned}
$$

which is equivalent to

$$
\int_{c_\theta^U}^{c^U(\theta')} p_\xi dF_\omega \geq [F_\omega(c^U(\theta')) - F_\omega(c_\theta^U)]c_\theta^U \tag{D.1}
$$

Similarly, for any θ choosing downward reserves, incentive compatibility requires:

$$
\begin{aligned}
\pi^D(\theta) &= p_\omega^F - c_\theta + (p_{D\theta}^K - I_\theta^D) + F_\omega(-p_D^X(\theta))(p_U^X(\theta) + c_\theta^D) \\
&\geq p_\omega^F - c_\theta + (p_D^K(\theta') - I_\theta^D) + F_\omega(-p_D^X(\theta'))(p_D^X(\theta') + c_\theta^D) \\
&= \pi^D(\theta') + I^D(\theta') - I_\theta^D + c(\theta') - c_\theta + F_\omega(-p_D^X(\theta'))(p_D^X(\theta') - p_D^X(\theta)) \\
&= \pi^D(\theta') + I^D(\theta') - I_\theta^D + [1 - F_\omega(c^D(\theta')](c(\theta') - c_\theta) \\
&+ F_\omega(c^D(\theta'))[(c(\theta') - c^D(\theta')) - (c_\theta - c_\theta^D)]
\end{aligned}
$$

which is equivalent to:

$$
\int_{c_\theta^D}^{c^D(\theta')} p_\xi dF_\omega \geq [F_\omega(c^D(\theta')) - F_\omega(c_\theta^D)]c_\theta^D \tag{D.2}
$$

□