This is the pre-published version published in
Energy Policy, available online at:
https://doi.org/10.1016/j.enpol.2019.110988.

A wholesale electricity market design *sans* missing money and price manipulation

C.K. Woo[a,*], I. Milstein[b], A. Tishler[c], J. Zarnikau[d]

[a] Department of Asian and Policy Studies, Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong

[b] Holon Institute of Technology, 52 Golomb St., Holon 58102, Israel

[c] Coller School of Management, Tel Aviv University, Tel Aviv 69978, Israel

[d] Department of Economics, The University of Texas at Austin, 2225 Speedway, Austin, TX 78712, USA

[*] Corresponding author

Email addresses: chiwoo@euduhk.hk (C.K. Woo); irenam@hit.ac.il (I. Milstein); ashert@tauex.tau.ac.il (A. Tishler); jayz@utexas.edu (J. Zarnikau)

JEL codes: D2, D43, L11, L94,

Keywords: Electricity market design, wholesale competition, reliability differentiation, tolling agreements, missing money, price manipulation

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Highlights

- Develop a new market design for wholesale electricity competition
- Document the design's solution for missing money and price manipulation
- Document the design's efficient planning, pricing and operation
- Document the design's practical implementation
- Recommend the design for reliability differentiation and market competition

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Abstract

Using reliability differentiation via tolling agreements with diverse heat rates and fuel types, we propose an efficient wholesale electricity market design under demand and supply uncertainty. Mainly based on North America's market experience, our proposed design adopts an independent system operator's (ISO's) existing practice of least-cost dispatch of heterogeneous generation units, real-time energy price determination and capacity rationing. It solves the missing money problem of inadequate incentive for generation investment, without requiring the ISO to operate centralized capacity auctions, make capacity payments, set high energy price caps, or subsidize market entry. It preempts independent power producers' price manipulation in the ISO's real-time market for energy, thus easing the ISO's burden of market monitoring. It suggests two-part pricing of end-use consumption of a load serving entity's retail customers, meaningfully linking the wholesale and retail markets. It is applicable to countries that have implemented wholesale competition or are in the process of doing so. Hence, its policy implication is that it should be considered in the ongoing debate of electricity reliability and market competition.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Abbreviations: AS, Ancillary services; CAISO, California Independent System Operator; CEC, California Energy Commission; CCGT, Combined cycle gas turbine; CT, Combustion turbine; DR, Demand response; DSS, Demand subscription service; ERCOT, Electricity Reliability Council of Texas; FSL, Firm service level; HR, Heat rate; IPP, Independent power producer; ISO, Independent system operator; kW, Kilowatt; kWh, Kilowatt hour; LLOE, Loss of load expectation; LLOP, Loss of load probability; LSE, Load serving entity; MMBtu, Million British thermal units; MW, Megawatt; MWh, Megawatt hour; NERC, North American Electricity Reliability Council; PJM, Pennsylvania-Jersey-Maryland; RAR, Resource adequacy requirement; RTM, Real-time market; VOLL, Value of lost load

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

1. Introduction

Motivated by the global trend of electricity market reform and deregulation (Sioshansi, 2013), we propose an efficient wholesale market design under demand and supply uncertainty. Mainly based on North America's market experience,[1] our proposed design's primary focus is electricity generation and wholesale market competition.[2] It uses reliability differentiation via tolling agreements with diverse heat rates and fuel types to solve the thorny problems of missing money and price manipulation. A real-world example of these two problems is Alberta's wholesale electricity market with an energy-only design (Brown and Olmstead, 2017).

The missing money problem occurs when a wholesale electricity market fails to provide adequate investment incentives for conventional generation units, including combined cycle gas turbines (CCGTs) and combustion turbines (CTs) (Joskow, 2013).[3] Exacerbating the missing money problem is the price reduction (*aka* merit order) effect of renewable generation like wind or solar that has zero fuel cost (Woo et al., 2011, 2013, 2014b, 2016a, 2017a, 2017b, 2018a; Zarnikau et al., 2014, 2019).[4]

The price manipulation problem occurs when independent power producers (IPPs) exercise their market power that can, even in the absence of a capacity shortage, cause

[1] Rich in details and pragmatic in nature, this paper draws heavily from the first author's publications based on research funded by several electric utilities and government agencies in North America, Israel and Hong Kong. As such, it has real-world applicability and relevance as detailed in Section 3 below.

[2] For details on transmission topics like power flow, congestion, line loss and network security, see Stoft (2002), Kumar et al. (2005) and Wood and Wollenberg (2012).

[3] Natural gas is the most popular fuel used by newly constructed generation units (Gal et al., 2017, 2019). Section 2 shows that our proposed market design can accommodate many heterogeneous generation technologies. We do not consider renewable generation's investment incentives which have been mainly driven by government intervention through generous tax credits, easy transmission access, above-market feed-in-tariffs, and legislated renewable portfolio standards (Alagappan et al., 2011; Zarnikau, 2011).

[4] A dramatic case in point is an electric grid's occasional inability to fully absorb the non-dispatchable generation from solar and wind resources. Consequently, negative prices are used to induce dispatchable generation (e.g., CCGT) owners to curtail their output so as to maintain the grid's real-time load-resource balance (https://energyathaas.wordpress.com/2017/04/24/is-the-duck-sinking/).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

abnormally high wholesale market prices (Wolfram, 1999; Borenstein et al., 2002; Woo et al., 2003a, 2006a). Price manipulation can also be the result of IPPs' gaming of a transmission system's congestion management (Stoft, 2002).[5] For example, wind generators in Texas formerly would submit exaggerated output schedules that would appear to result in transmission congestion. Then, they would get paid to resolve the fake transmission congestion.[6] To be fair, high market prices can occur *sans* market power abuse because of IPPs' rational behavior (Milstein and Tishler, 2012).

Our proposed design is attractive because it does not require the various measures designed to enhance IPPs' investment incentives. Examples of such measures include centralized capacity auctions, capacity payments, high energy price caps, and subsidized market entry (Spees et al., 2013; Cramton, 2017; Bublitz et al., 2019). It also preempts IPPs' market abuse in an independent system operator's (ISO's) real-time market (RTM) for energy, thus easing the ISO's burden of market monitoring.[7]

Our proposed design stems from reliability differentiation of wholesale transmission service (Woo et al., 1998) and retail end-use service (Chao and Wilson, 1987; Seeto et al., 1997; Woo et al., 2014a). In particular, it relies on the concept of demand subscription service (DSS) with fixed cost recovery (Woo, 1990, 1991). Under DSS, an end-use customer of an integrated utility must subscribe, before actual consumption, a firm service level (FSL in kW) that cannot be curtailed by the utility during a generation capacity shortage. The customer may, however, still see service

[5] An overview of Texas's transmission congestion management is available at: http://www.ercot.com/content/wcm/training_courses/109518/Nodal_101.pdf.
[6] An example of a wind generator's gain from overscheduling is the negative market price described in footnote 4 times the MWh difference between the scheduled and actual generation.
[7] A description of the California Independent System Operator's market monitoring is available at http://www.caiso.com/market/Pages/MarketMonitoring/Overview.aspx.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

interruptions due to local transmission and distribution failures. The utility's optimal generation capacity level is the total MW size of all FSL subscriptions plus a reserve margin to account for random plant outages, greatly simplifying the problem documented by Hobbs (1995) of utility resource planning under demand and supply uncertainty.

A wholesale electricity market typically comprises an ISO, IPPs, load serving entities (LSEs), and a regulated transmission company (Woo et al., 2003a).[8] Under open transmission access (Lusztig et al., 2006), wholesale electricity trading may occur bilaterally or though the ISO's centralized markets (Woo et al., 2013, 2017a, 2017b, 2018a; Zarnikau et al., 2014, 2019).[9] We use the DSS concept to develop a new wholesale market design with the following key elements: (a) an ISO's RTM price determination and capacity rationing; (b) LSEs' two-part pricing of end-use consumption of their retail customers; (c) LSEs' optimal procurement plans for tolling agreements; and (d) LSEs' execution of the plans in (c) via decentralized procurement auctions. While these elements are not new, their meaningful integration under our proposed design is.

Presently absent in North America's wholesale electricity markets is our proposed design's inducement of a LSE to self-reveal its reliability preference, enabled by the requirement that ***all*** LSEs under an ISO's jurisdiction must procure tolling agreements. This must-procure requirement is unrestrictive in practice because the total MW size of a LSE's procured agreements can range from zero MW to the peak MW unmet by the resources already in place (e.g., previously signed take-or-pay forward contracts and

[8] A LSE may be a local distribution company created by the divesture of a formerly integrated utility or a retail service provider emerged after the introduction of retail competition. For states (e.g., Oregon and Washington in the U.S.) and provinces (e.g., British Columbia and Quebec in Canada) that do not have an ISO, the regulated transmission company is the system operator.
[9] A list of California's centralized markets and products is available at: http://www.caiso.com/market/Pages/MarketProcesses.aspx A similar list for Texas is available at: http://www.ercot.com/mktinfo

wind power purchase agreements). As a LSE preferring higher reliability tends to procure more MWs, the ISO can use the LSEs' procurement outcomes for efficient capacity rationing.[10]

Our proposed design is practical. Specifically, it uses an ISO's existing practice of (a) capacity rationing during a grid's critical hours;[11] and (b) RTM operation and price determination, like those described by Woo et al. (2018a) for the California Independent System Operator (CAISO) and Zarnikau et al. (2014, 2019) for the Electric Reliability Council of Texas (ERCOT).[12] It also uses a regulated LSE's Hopkinson tariff design (Seeto et al., 1997) for pricing retail consumption and auction process (Woo et al., 2004, 2016b) for procuring tolling agreements. Its practicality is further underscored by what the ISO does ***not*** use, including capacity payments and price caps (Milstein and Tishler, 2019), subsidized market entry (Brown, 2018a), centralized capacity auctions (Brown, 2018b), and cost auditing of generation units (Munoz et al., 2018). As a result, it is applicable to countries that have implemented wholesale market competition (e.g., the U.S. and Canada in North America; Brazil and Chile in South America; Britain, Germany,

[10] If a LSE's retail customers are mainly firms with onsite backup generation, the LSE likely prefers low reliability (Woo and Pupp, 1992). In contrast, if a LSE's retail customers are mostly high-income households with staying home members, the LSE likely prefers high reliability (Woo et al., 2014c and references thereof). As a result, a LSE's reliability preference may initially come from a general understanding of retail customers.

Section 2.4.1 suggests a LSE's two-part retail pricing that has a demand charge ($/kW-month) for FSL subscription and a real time energy charge ($/kWh) for energy consumption. Appendix 2 shows that the total MW of all FSL subscriptions helps set the total MW target of a LSE's procurement auction.

[11] A description of California's system alerts and emergency actions is available at https://www.caiso.com/Documents/SystemAlertsWarningsandEmergenciesFactSheet.pdf

[12] We focus on RTM prices even though CAISO and ERCOT have hourly day-ahead market prices set on day *d*-1 for energy delivered on day *d*. As these day-ahead prices are based on the day-ahead forecast of system conditions (Woo et al., 2017a; Zarnikau et al., 2019), they are not the market clearing prices for achieving an electric grid's real-time load-resource balances.

Both CAISO and ERCOT use security-constrained economic dispatch algorithms to manage energy production, system power balance and network congestion, yielding 5-minute nodal prices (Woo et al., 2018a; Zarnikau et al., 2014, 2019). Based on the theory of locational marginal pricing (Stoft, 2002), these nodal prices are the supply bid prices of the last dispatched units at the grid's numerous nodes. If these supply bid prices are excessively high due to IPPs' market abuse under lax market monitoring, they can cause inflated nodal prices.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Spain and Norway in Europe; Australia, New Zealand and Singapore in Asia Pacific) or are in the process of doing so (e.g., China, Israel, Japan and Korea in Asia).

Our main contribution is our newly developed market design's reliability differentiation via tolling agreements based on many ***heterogeneous*** generation technologies with ***diverse*** heat rates and fuel types, unlike the single- and two-technology representations typically assumed in a Cournot market analysis (e.g., Milstein and Tishler, 2012, 2019; Gal et al., 2017, 2019). It underscores our methodological innovation of using tolling agreements to design an efficient wholesale electricity market under demand and supply uncertainty *sans* the twin problems of missing money and price manipulation.

To place our contribution in the context of electricity reliability and market competition, our proposed design is an extension of mandatory reservation for firm transmission service under the *pro forma* tariff in Order 888 of the U.S. Federal Energy Regulatory Commission (Woo et al., 1998). It is a variation of the design proposal of Chao (2012) that uses demand-side forward contracts for priority service developed by Chao and Wilson (1987). Unlike the design proposal of Chao and Wilson (2004) that uses capacity call options with demand-based strike prices, it relies on heterogeneous tolling agreements with supply-based strike prices amenable to an ISO's least-cost dispatch, leading to a cost-based RTM price determination that resembles those of South America's designs described by Munoz et al. (2018). Finally, it complements the analysis of Joskow and Tirole (2007), bridging the gap among various stakeholders, including: economists focused on competitive market mechanisms; engineers focused on an electric grid's physical attributes; industry practitioners focused on a market design's practicality; financial analysts focused on investment risks and return; LSEs focused on how to best serve their retail customers; retail customers focused on price stability, price reasonableness and service choice; IPPs focused on their financial performance; and regulators and policy makers focused on reliable electricity service at competitive prices, incentive compatibility and customer choice.

The rest of the paper proceeds as follows. Section 2 formulates our proposed market design and presents our key findings, which are further discussed in Section 3. Section 4 concludes by summarizing our proposed design's attractive properties, yielding the policy implication that the design should be considered in the ongoing debate of electricity reliability and market competition (Spees et al., 2013; Cramton 2017; Coester et al., 2018; Conejo and Sioshansi, 2018; Newbery et al., 2018; Bublitz et al., 2019).

2. Materials and methods

2.1 Overview

Intended for a general audience, this section is an overview of our proposed design to introduce the key concepts used in the next three sections. These concepts are tolling agreements, procurement auction, reliability differentiation, and an ISO's least-cost dispatch, RTM price determination and capacity rationing.

A tolling agreement is often based on a thermal generation unit (Eydeland and Wolyniec, 2003; Deng and Oren, 2006), a bilateral contract useful for managing a LSE's procurement cost and risk (Woo et al., 2006b).[13] It has terms and conditions that govern the underlying generation unit's forced outage rate and maintenance requirement.[14] Hence, the LSE's optimal procurement plan described in Section 2.4.1 accounts for the

[13] The bilateral market for tolling agreements is not a centralized capacity market operated by an ISO like those in New York, New England and Pennsylvania-Jersey-Maryland (PJM). Hence, our paper is not about the merit of a centralized capacity market.
[14] For the sake of clarity without any loss in generality, we assume that the unit's owner is responsible for maintenance, whose cost is included in the agreement's capacity price.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

supply uncertainty due to the unit's forced outage rate. Under the assumption made in Section 2.2 that the ISO controls the unit's production, the unit's scheduled maintenance likely occurs in the mild-weather spring months of April and May when the grid has surplus capacity because of low hourly system demands.[15]

After paying the agreement's upfront capacity price D (\$/MW-year), a buying LSE has the right, but not the obligation, to obtain electricity from a selling IPP within the agreement's contract duration at the strike price defined below:[16]

$$\$C\text{/MWh} = \text{Heat rate } HR \text{ (MMBtu/MWh)} \times \text{Fuel price } F \text{ (\$/MMBtu)}. \tag{1}$$

The LSE exercises the right when the hourly spot market price P (\$/MWh) exceeds C, earning an *ex post* per MWh variable profit of $\max(P – C, 0)$ that equals the per MWh cost saving from not buying electricity at high spot market prices (Woo et al., 2016b).

We use C in equation (1) to characterize heterogeneous tolling agreements.[17] For example, a CCGT-based agreement's C is lower than a CT-based agreement's because a CCGT's HR is lower than a CT's. This characterization is general, equally applicable to generation technologies with different heat rates and fuel types (e.g., natural-gas-fired vs. coal-fired generation).

---

[15] California's maintenance of thermal generation plants mainly occurs in these two months, an outcome reinforced by the state's inexpensive (< \$20/MWh) hydro power import during the Pacific Northwest's spring runoff (Woo et al., 2013, 2017b).

[16] The variable HR measures a generation unit's efficiency in converting fuel (e.g., natural gas) to electricity. Both HR and F in equation (1) are stipulated in the bilateral agreement signed by the LSE and the IPP. While outside the purview of an ISO, they may be subject to prudence review of a local regulator (e.g., a state public utility commission) if the LSE is a regulated local distribution company (Woo et al., 2006c).

For expositional simplicity, equation (1) intentionally ignores the per MWh costs for variable O&M, startup and ramping. Including these additional cost terms arithmetically complicates our market design analysis without the benefit of additional insights.

[17] Equation (1) encompasses a demand response (DR) contract, which allows a LSE, after paying an upfront capacity price, to have the right but not the obligation, to obtain load reduction at \$$C$/MWh from a DR supplier. Hence, our proposed design can incorporate DR contracts such as California's curtailable rate option (Moore et al., 2010) and hybrid capacity option (Woo et al., 2018b).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

While *C* in equation (1) may fluctuate randomly (Gal et al., 2017, 2019), it is known with certainty on the day when the LSE decides whether to exercise the right to obtain electricity from the IPP. In short, a tolling agreement is an exotic option, comprising hourly capacity call options with daily varying strike prices (Deng et al., 2001). The missing money problem arises when a tolling agreement's expected variable profit cannot cover the underlying generation unit's returns on and of investment (Woo et al., 2016b).

A tolling agreement's capacity price yields a stable revenue stream for an IPP, an important consideration in the IPP's ability to obtain project financing at reasonable terms (Stern, 1998). It decreases with *C* because the installation cost of a generation unit with a relatively low per MWh fuel cost (e.g., a CCGT) exceeds that of a generation unit with a relatively high per MWh fuel cost (e.g., a CT).

Our proposed design assumes competitively determined capacity prices for tolling agreements. Absent an active market for tolling agreements, such prices can come from a LSE's procurement auction (Laffont and Tirole, 1993), as currently done in California (Woo et al., 2016b). Based on a winning IPP's revealed preference, a signed agreement's capacity price presumably embodies adequate investment incentive that includes the returns on and of investment.

Enabling our proposed design's reliability differentiation is the ISO's requirement that ***all*** LSEs under its jurisdiction must procure tolling agreements. This must-procure requirement is unrestrictive in practice because a LSE preferring higher reliability can procure more MWs. Further, a LSE that owns generation units or has previously signed

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

power contracts can choose to procure between zero MW and the peak MW load unmet by the resources already in place.

After confirming all LSEs' procured tolling agreements, the ISO uses a least-cost dispatch of the available generation units to serve each LSE's time-dependent demands (Chao, 1983; Stoft, 2002), yielding a RTM price solely based on the last dispatched unit's per MWh fuel cost. As shown by Figure 1 below, the ISO's RTM price determination is not bid-based, thus preempting IPPs' RTM price manipulation (Munoz et al., 2018).

Following the DSS's load curtailment strategy (Woo 1990, p.71), the ISO's capacity rationing scheme under our proposed design is as follows. When the grid has a capacity surplus, ***all*** LSEs' demands are fully met. When the grid has a capacity shortage, each LSE receives its total procured capacity that is available during the shortage. Thus, the scheme precludes the inefficient outcome of load curtailment for ***some*** LSEs when the grid has a capacity surplus. It also implies that each LSE sees a load-resource balance constraint in a shortage hour.

The ISO's capacity rationing is *ex post* efficient when it results in equal marginal benefits of electricity consumption for all LSEs (Woo et al., 2008). A sufficient condition for *ex post* efficiency is that LSEs' marginal benefits of electricity consumption are weakly separable, exhibiting similar time dependence and weather sensitivity (Woo, 1990, pp. 76-77). Satisfying this condition are the popular functional forms (e.g., double-log and linear) used in electricity demand estimation (Woo, 1993; Woo et al., 2018c). Hence, the ISO's capacity rationing can empirically be as *ex post* efficient as real-time pricing proposed by Bohn et al. (1984).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Matching the time sequence of the LSEs' resource procurement and the ISO's RTM operation,[18] we use three steps to demonstrate our proposed market design's economic efficiency and adequate incentive for generation investment. In Step 1, the LSEs optimally procure tolling agreements before the realization of hourly RTM prices and weather conditions. In Step 2, the LSEs inform the ISO of their real-time electricity demands under the realized weather condition. In Step 3, the ISO uses the LSEs' procured agreements and real-time electricity demands for its least-cost dispatch, RTM price determination and capacity rationing. We analyze these steps recursively in Sections 2.2 to 2.4 below.

2.2 Step 3: The ISO's least-cost dispatch, RTM price determination and capacity rationing

We assume N LSEs that have agreed to delegate the right to the ISO for obtaining electricity at their procured tolling agreements' strike prices.[19] The ISO performs a least-cost dispatch based on the LSEs' procured tolling agreements and real-time electricity demands. To do so, the ISO first confirms: (a) LSE n's procured MWs by tolling agreement type are $\{K_{jn}\}$ for $j = 1, \ldots, J$ and $n = 1, \ldots, N$; and (b) the hourly per MWh fuel cost of K_{jn} is C_{jh} in hour $h = 1, \ldots, H = 8{,}760$ = number of hours in a calendar year.[20]

[18] This time sequence follows those of CAISO (Woo et al., 2018a) and ERCOT (Zarnikau et al., 2019).

[19] The size of N is relatively small for California where retail competition is limited. In contrast, N is relatively large for Texas, a state with robust retail competition (Distributed Energy Financial Group, 2015). We assume that these LSEs aim to best serve their customers for the following reasons. First, if the retail market is highly competitive, unregulated LSEs compete rigorously to attract and retain customers. Second, if the retail market is dominated by a few unregulated LSEs, regulatory oversight and sanction restrain these LSEs' price gouging and other non-competitive practices. Finally, regulated LSEs are mandated to efficiently serve their customers at just and reasonable rates.

[20] Setting H at 8,760 reflects our focus of an annual analysis, which can be readily changed by modifying H's size. For example, a 10-year analysis would set $H = 87{,}600$ and entail discounted prices and costs to reflect the time value of money. We adopt the 60-minute time scale mainly for notational simplicity. The analysis presented below is equally valid for the 5-minute time-scale, involving arithmetic changes in the unit of measurement and time-interval definition. For example, the energy embodied in a 5-minute demand of Z MW is $(Z / 12)$ MWh. Similarly, H becomes $12 \times 8760 = 105{,}120$ for the 5-minute time scale.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

As there is no size limit for J, the procured agreements of ***all*** LSEs can encompass many heterogeneous generation technologies with diverse heat rates and fuel types. Moreover, the ISO's confirmation uses the signed tolling agreements submitted by the LSEs, not a cost audit of individual generation units. This makes sense because a tolling agreement is a commercial contract, with clearly stated terms and conditions that govern and enforce the buyer's and the seller's contractual obligations.

The ISO cost ranks the agreements procured by ***all*** LSEs, resulting in a system merit order of 1 to J that matches $C_{1h} < \ldots < C_{Jh}$.[21] This merit order is assumed to be time-invariant for two reasons. First, if all underlying generation units burn the same fuel (e.g., natural gas), the merit order reflects these units' engineering-based heat rates of $HR_1 < \ldots < HR_J$. Second, nuclear and coal-fired units tend to have relatively low per MWh fuel costs and be dispatched before natural-gas-fired units (Woo et al., 2014b).

Let $Q_h = Q(P_h, h, t_h)$ denote the grid's aggregate electricity demand of N LSEs at hourly price P_h and realized weather index t_h. An example of a weather index is the degree-hour commonly used in electricity demand studies, conveniently defined herein as $t_h \equiv$ |hourly temperature – 65°F|.[22] When the hourly temperature rises above or falls

[21] This merit order ignores CO_2 emissions, transmission congestion, and line loss and the ISO's need for operating reserves (i.e., ancillary services (AS)) for real-time grid operation. Including such details does not qualitatively alter our key findings. To wit, the merit order may be revised to reflect the per MWh costs that contain the cost effects of CO_2 emissions (Woo et al., 2018a) and transmission congestion and line loss (Bohn et al., 1984). Similarly, the ISO's AS acquisition may come from the ISO's day-ahead assessment of system conditions, which is then used in the ISO's real-time security-constrained economic dispatch (Zarnikau et al., 2019). Finally, the merit order may need to account for LSE n's existing tolling agreements and dispatchable generation units.

What matters here is that the resulting per MWh costs can still be used for the profit calculation based on equation (4) below. Consequently, the analysis in Sections 2.3 and 2.4 do not qualitatively depend on the computational nuances of the per MWh costs.

[22] Changing our choice of 65°F does not alter our analysis in the rest of this paper.

below 65°F, t_h increases, raising the grid's aggregate electricity demand through the LSEs' weather-sensitive loads.[23]

We define $Q_h = \Sigma_n\, q(P_h, h, t_h, n)$, where $q(P_h, h, t_h, n)$ = LSE n's hourly electricity demand that does not *a priori* restrict t_h to be an additive or a multiplicative term. In contrast, a Cournot market analysis typically assumes that t_h is an additive term in the linear market demand function (e.g., Milstein and Tishler, 2012, 2019; Gal et al., 2017, 2019). The Step-2 analysis in the next section shows that $q(\bullet)$ is decreasing in P_h but increasing in t_h, in line with the empirics reported in the vast literature of electricity demand estimation.[24]

As a generation unit may become unavailable in hour h due to an unexpected equipment failure, we use $K_h = \Sigma_n\, \phi(\{K_{jn}\},h)$ to denote the grid's total capacity made available in hour h by N LSEs' procurements of $\{K_{jn}\}$. Thus, $\phi(\bullet)$ mimics a production function, transforming the procured MWs into available MWs. It is increasing in K_{jn} with $\partial\phi(\bullet)/\partial K_{jn} = \beta_{jh} = 1$ if K_{jn} is available in hour h; 0 otherwise.

Subject to the grid's capacity constraint of $Q_h \leq K_h$, the ISO's least-cost dispatch of generation units follows the system merit order so that the last dispatched unit j^* obeys the condition of $C_{1h} < \ldots < C_{(j^*-1)h} < C_{j^*h} < C_{(j^*+1)h} < \ldots < C_{Jh}$. Generation units with per MWh fuel costs above C_{j^*h} are not dispatched in hour h and therefore have zero output.

The ISO sets the RTM's market-clearing price (\$/MWh) for hour h at the last dispatched unit's per MWh fuel cost:

[23] For example, California is a summer peaking state whose high hourly demands occur on hot afternoons due to high air conditioning loads (Woo et al., 2014b). The same can be said for other summer peaking states like Arizona, Nevada, New Mexico and Texas. In contrast, winter-peaking states in the Pacific Northwest have high hourly demands that occur on cold winter evenings due to high space heating loads (Woo et al., 2013).

[24] For literature reviews of electricity demand estimation, see Faruqui and Sergici (2010), Newsham and Bowker (2010) and Woo et al. (2018c).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

$$P_h = C_{j*h}. \qquad (2)$$

The ISO's RTM price determination given by equation (2) replaces a Cournot electricity market's price outcome (Milstein and Tishler, 2012, 2019; Gal et al., 2017, 2019). IPPs cannot manipulate P_h in equation (2), chiefly because P_h is set *sans* IPPs' supply bidding (Munoz et al., 2018). Because P_h's volatility is the same as C_{j*h}'s fuel-price-driven volatility, it is far less than the actual RTM price volatilities observed in states like California (Woo et al., 2018a) and Texas (Zarnikau et al., 2019).[25]

The ISO's capacity rationing scheme operates as follows. If $Q_h \leq K_h$, LSE n's *ex post* hourly consumption is the fully met $q(\bullet)$ for $t_h < \tau_h$ = weather index at which $Q_h = K_h$;[26] otherwise, it is $K_{nh} = \phi(\bullet)$ = LSE n's *ex post* capacity available. As a result, the condition of $K_{nh} = \phi(\bullet)$ is LSE n's load-resource balance constraint in a shortage hour.

The ISO's capacity rationing scheme implies that when the grid has an hourly capacity surplus, LSE n's *ex post* consumption may differ from *ex post* capacity available. Should the ISO use LSE n's excess capacity to meet other LSEs' demands, it would provide a profit refund to LSE n shown by equation (3) below.

The grid's loss-of-load probability in hour h is $LOLP_h = 1 - G(t_h = \tau_h)$, where $G(t_h)$ = cumulative distribution function of $t_h \in (0, T$ = maximum degree hour). The grid's annual loss-of-load expectation is $LOLE$ = 8,760 hours × Σ_h $LOLP_h$. As τ_h increases with K_h that depends on the LSEs' procurements, the grid's $LOLE$ is endogenously determined

[25] Gal et al. (2017) document the fuel price volatilities of coal, oil and natural gas, which are much lower than the wholesale electricity price volatilities observed in North America.

[26] The weather index threshold τ_h varies hourly because Q_h is time-dependent. To wit, an electric grid's hourly demands during the 00:00 to 06:00 period are typically lower than those during the 06:00 to 24:00 period of a given day.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

under our proposed market design, unlike the administratively set *LOLE* criterion of 1-day-in-ten-years for North America's electricity industry.[27]

Figure 1 illustrates the ISO's RTM price determination and capacity rationing. The solid green line portrays a hypothetical grid's generation stack of K_{1h} MW of CCGTs and K_{2h} MW of CTs available in hour h. The per MWh fuel costs for the CCGTs and CTs are C_{1h} and C_{2h} respectively. The red dashed lines are the grid's demands in hour h by weather condition. The grid's RTM price is $P_h = C_{1h}$ under mild weather and $P_h = C_{2h}$ under extreme weather.[28] Rather than allowing P_h to rise when the weather condition is extreme, the ISO uses capacity rationing in the shortage hour to resolve the difference between the grid's aggregate Q_h at $P_h = C_{2h}$ and total available capacity at $K_h = K_{1h} + K_{2h}$.

Our proposed design's financial settlement is as follows. LSE n pays P_h for each MWh bought from the ISO's RTM. The ISO's total RTM revenue is the sum of the energy payments made by all LSEs. The ISO uses the total RTM revenue to pay the fuel costs incurred by the dispatched generation units. Unless $P_h = C_{1h}$ = minimum RTM price for all hours, the ISO's total RTM revenue exceeds total fuel cost payment, resulting in a strictly positive operating surplus. The ISO refunds the surplus according to LSE n's *ex post* profit given below:

$$V_{hn} = (P_h - C_{1h})\,\beta_{1h}\,K_{1n} + \ldots + (P_h - C_{(j^*\text{-}1)h})\,\beta_{(j^*\text{-}1)h}\,K_{(j^*\text{-}1)n}. \tag{3}$$

[27] Reliability standards of the North American Reliability Council (NERC) are available at https://www.nerc.com/pa/Stand/Reliability%20Standards%20Complete%20Set/RSCompleteSet.pdf

[28] The market clearing price P_h in Figure 1 is between C_{1h} and C_{2h} when the grid's hourly aggregate demand of Q_h MW ***exactly*** equals the total CCGT capacity of K_1 MW. This is an ***extremely rare*** technicality that we decide to ignore.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Equation (3) reflects that at $P_h = C_{j*h}$, (K_{1n}, …, $K_{(j*-1)n}$) are profitable and K_{j*n} earns zero profit. Owing to the fuel cost payments and profit refunds, the ISO's RTM operation breaks even with certainty.[29]

A general representation of equation (3) is:

$$V_{hn} = \max(P_h - C_{1h}, 0)\, \beta_{1h}\, K_{1n} + \ldots + \max(P_h - C_{Jh}, 0)\, \beta_{Jh}\, K_{Jn}, \quad (4)$$

implying that the marginal profit effect of K_{jn} is:

$$\partial V_{hn} / \partial K_{jn} = \max(P_h - C_{jh}, 0)\, \beta_{jh}. \quad (5)$$

2.3 Step 2: A LSE's real-time energy demand assessment

We assume LSE n passes α_n% of V_{hn} to its retail customers. Hence, $\alpha_n V_{hn}$ may be seen as a "dividend payout" from LSE n to its retail customers in hour h. The following cases suggest α_n likely near 100%: (1) LSE n is a non-profit municipal utility; (2) LSE n is a local distribution company subject to the cost of service regulation; and (3) LSE n is an unregulated company operating in a highly competitive retail market. For the sake of generality, we allow α_n to lie between 0% and 100%.

The real-time energy demand curve $q(P_h, h, t_h, n)$ comes from LSE n's benefit maximizing customers. Communicated to the ISO shortly (e.g., within 60 minutes) before the real-time dispatch, $q(\bullet)$ indicates the MWh amounts that LSE n is willing to buy at different RTM price levels in hour h when the already known weather index is t_h.[30] As Figure 1 shows that LSE n's MWh demanded cannot be fully met at $P_h = C_{j*h}$ in a

[29] In the U.S., an ISO uses cost of service ratemaking (e.g., access and customer charges based on embedded costs) to recover its non-electricity costs for administration, management, market surveillance, billing, …, etc. (Lusztig et al., 2006).

[30] An hourly electricity demand tends to be highly price inelastic because of its short run nature. The quantity demanded is mainly driven by its time dependence and weather sensitivity. LSE n's assessment of $q(\bullet)$ can be based on an estimation of electricity demand under time-of-use pricing, as exemplified by studies reviewed by Faruqui and Sergici (2010) and Newsham and Bowker (2010) for summer peaking LSEs and Woo et al. (2017c) for winter peaking LSEs.

shortage hour, the ISO's capacity rationing scheme enters LSE n's optimal procurement plan, see Section 2.4.1 below.

To derive $q(\bullet)$, define the hourly maximum benefit from electricity consumption by LSE n's customers *sans* the ISO's curtailment (Woo, 1990, p.72):

$$W_{hn} = [\max_{q_h} \int p(q, h, t_h, n)\, dq - P_h\, q_h] + \alpha_n V_{hn}, \quad (6.a)$$

$$= U_{hn} + \alpha_n V_{hn}, \quad (6.b)$$

where $p(q, h, t_h, n)$ = marginal benefit of consuming q MWh of electricity for $q \in (0, \infty)$ and U_{hn} = square-bracket term in equation (6.a). Equation (6.a) states that the maximum benefit W_{hn} has two parts: (1) U_{hn} from consuming q_h MWh of electricity; and (2) $\alpha_n V_{hn}$ received from LSE n.

Based on Woo (1990), we assume $\partial p(\bullet)/\partial q < 0$ and $\partial p(\bullet)/\partial t_h > 0$, reflecting that LSE n's marginal benefit decreases with electricity consumption but increases with the weather index. As LSE n's $q(\bullet) > 0$, it satisfies W_{hn}'s first order condition of $p(\bullet) = P_h$, lending support to the real-time marginal cost pricing rule (Chao, 1983; Bohn et al., 1984; Stoft, 2002; Woo et al., 2008). We use comparative statics to verify that $q(\bullet)$ is decreasing in P_h but increasing in t_h.

2.4 Step 1: A LSE's optimal procurement of tolling agreements

In Step 1, LSE n procures tolling agreements at competitively determined capacity prices. LSE n's procurement process answers the questions of "what to buy" and "how to buy", a variation of theme of a LSE's management of procurement cost and risk

(Woo et al., 2004).[31] We assume LSE n is risk-neutral. We later discuss the effect of risk aversion on LSE n's procurement plan in Section 4.

2.4.1 What to buy

Answering the "what to buy" question requires LSE n to develop an optimal procurement plan. Suppose the capacity prices are $\{D_j\}$with $D_1 > \ldots > D_J$, reflecting that agreements with lower per MWh fuel cost forecasts have higher capacity prices than those with higher per MWh fuel costs forecasts.[32]

We assume that LSE n aims to best serve its customers, reflecting the policy mandate of a regulated LSE and the retail competition faced by an unregulated LSE. Hence, LSE n's objective function that includes the expectation of W_{hn} and retained profit of $(1 - \alpha_n)\, V_{hn}$ is:

$$\Omega_n = \Sigma_h \,\mathrm{E}[W_{hn} + (1 - \alpha_n)\, V_{hn}] - M_n, \tag{7.a}$$

$$= \Sigma_h \,\mathrm{E}(U_{hn} + V_{hn}) - M_n, \tag{7.b}$$

where M_n = LSE n's total capacity payment = $\Sigma_j D_j K_{jn}$. Equation (7.b) is the result of recognizing equation (6.b) that states $W_{hn} = U_{hn} + \alpha_n V_{hn}$.

To find its optimal procurement plan, LSE n chooses $\{K_{jn}\}$ to maximize Ω_n, subject to the ISO's capacity rationing scheme. To solve LSE n's maximization problem, we recall K_{nh} is LSE n's *ex post* consumption in a shortage hour. We also note that K_{nh} does not vary with $t_h \in (\tau_h, T)$ because it is capped by the ISO's capacity rationing scheme. As a result, K_{nh}'s likelihood of occurrence is $LOLP_h$. Further, $\mathrm{E}[\phi(\bullet) - K_{nh}] = 0$ is LSE n's constraint of expected load-resource balance in a shortage hour.

[31] We do not consider the auction's timing and frequency. The "when to buy" and "how often" questions are implementation issues well beyond the intent and scope of our market design analysis.
[32] These per MWh fuel cost forecasts can be based on the forward fuel prices in LSE n's planning period (Sreedharan et al., 2012).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Let S be the set of shortage hours. The Lagrangian with hourly multipliers $\{\lambda_{hn}\}$ is:

$$L = \Omega_n + \Sigma_{h\in S}\, LOLP_h\, \lambda_{hn}\, \mathrm{E}[\phi(\bullet) - K_{nh}]. \tag{8}$$

The Kuhn-Tucker conditions for LSE n's optimal procurement of $K_{jn} \geq 0$ are:

$$\partial L/\partial K_{jn} = \partial \Omega_n/\partial K_{jn} + X_n + \Sigma_{h\in S}\, LOLP_h\, \lambda_{hn}\, \mathrm{E}[\partial\phi(\bullet)/\partial K_{jn}] + Y_n - D_j \leq 0; \tag{9.a}$$

$$K_{jn}\, \partial L/\partial K_{jn} = 0; \tag{9.b}$$

$$\partial L/\partial \lambda_{hn} = LOLP_h\, \mathrm{E}[\phi(\bullet) - K_{nh}] \leq 0; \tag{9.c}$$

$$\lambda_{hn}\, \partial L/\partial \lambda_{hn} = 0. \tag{9.d}$$

In equation (9.a), X_n and Y_n are terms related to the marginal effect of K_{jn} on the endogenously determined τ_h because an increase in K_{jn} raises the capacity available at the system level. Specifically, X_n is the sum of: (a) $\Sigma_h (U_{hn} + V_{hn})\, g(\tau_h)\, \partial\tau_h/\partial K_{jn}$ evaluated at $q(\bullet)$ and (b) $\Sigma_h -(U_{hn} + V_{hn})\, g(\tau_h)\, \partial\tau_h/\partial K_{jn}$ evaluated at K_{nh}, where $g(t_h)$ = density function of t_h. As $q(\bullet) = K_{nh}$ at $t_h = \tau_h$, $X_n = 0$. Further, $Y_n = - \Sigma_{h\in S}\, \lambda_{hn}\, \mathrm{E}[\phi(\bullet) - K_{nh}]\, g(\tau_h)\, \partial\tau_h/\partial K_{jn} = 0$ because $\phi(\bullet) = K_{nh}$ in a shortage hour,

Equations (9.a) to (9.d) yield ***market-based*** findings that match the efficient pricing and planning rules in Chao (1983) and Woo (1990) for social welfare maximization, affirming our proposed design's incentive compatibility (Laffont and Triole, 1993). To establish this claim, we first define $A_j = \partial\Omega_n/\partial K_{jn} = \Sigma_h\, \partial \mathrm{E}(V_{hn})/\partial K_{jn} = \Sigma_h\, \mathrm{E}[\max(P_h - C_{jh}, 0)\, \beta_{jh}]$ for all $n \in (1, N)$, which is the market-based profit expectation (\$/MW-year) of an additional MW of capacity provided through tolling agreement type j.

We next define $B_{jn} = \Sigma_{h\in S}\, LOLP_h\, \lambda_{hn}\, \mathrm{E}[\partial\phi(\bullet)/\partial K_{jn}]$, which is LSE n's per MW-year expected marginal capacity value for tolling agreement type j. An estimate for B_{jn} is $(\Sigma_{h\in S}\, LOLP_h \times VOLL_{hn} \times f_j)$, where $VOLL_{hn} \approx \lambda_{hn}$ is LSE n's value of lost load in shortage hour h; and $f_j \approx \mathrm{E}[\partial\phi(\bullet)/\partial K_{jn}]$ is the availability factor of agreement type j based on a

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

generation technology's forced outage rate (Chao, 1983, p.184; Woo, 1990, p.75). This B_{jn} estimate makes sense because a 1-MW increase in ***firm*** capacity yields a reliability benefit of reducing LSE n's annual expected outage cost by an amount equal to ($\Sigma_{h\in S}$ $LOLP_h \times VOLL_{hn}$). Since the availability factor is $f_j < 1$, the incremental MW's reliability benefit declines accordingly.

Using the definitions of A_j and B_{jn}, we now state the key findings associated with LSE n's optimal procurement plan. First, if $(A_j + B_{jn}) < D_j$, $K_{jn} = 0$, reflecting that the per MW benefit $(A_j + B_{jn})$ is less than the per MW cost D_j (Chao 1983, p.184).

Second, suppose $K_{jn} > 0$ and $(A_j + B_{jn}) = D_j$. While the expected profit A_j is less than the capacity price D_j, LSE n is willing to pay a premium of B_{jn} for procuring K_{jn}. This finding justifies California's use of $(D_j - A_j)$ to estimate the marginal capacity value of a new generation unit in a LSE's procurement plan (Sreedharan et al., 2012, p.117).

Third, $(D_j - D_{j'}) = (A_j - A_{j'}) + (\Sigma_{h\in S}\, LOLP_h \times VOLL_{hn})\,(f_j - f_{j'})$ for $K_{jn} > 0$ and $K_{j'n} > 0$. When $f_j = f_{j'}$, $(D_j - D_{j'}) = (A_j - A_{j'})$. Hence, the capacity price difference between two chosen tolling agreements with identical availability should equal their market-based profit difference. This profit-based finding mirrors the cost-based result of Chao (1983, p.183): the capacity cost difference should equal the expected fuel cost difference between two chosen generation technologies.

Fourth, LSE n cannot break even because $A_j < D_j$ for $K_{jn} > 0$. This finding parallels the negative profit remark of Chao (1983, p.187). In an effort to break even, LSE n calculates a monthly demand charge applicable to a retail customer's FSL

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

subscription (Seeto et al., 1997).[33] Hence, LSE *n*'s two-part retail pricing uses a Hopkinson tariff design, comprising a monthly demand charge ($/kW-month) for FSL subscription and a real-time charge ($/kWh) for energy consumption.

We recognize that DSS may not have immediate acceptance by LSE *n*'s retail customers. Hence, the two-part pricing's implementation uses an opt-in approach under which a retail customer's default monthly FSL subscriptions are the customer's monthly peak demands. If the customer desires curtailable service in exchange for a bill reduction, it can make monthly FSL subscriptions that are below its monthly peak demands (Moore et al., 2010; Woo et al., 2014a).

Finally, we verify that the optimal $\{K_{jn}\}$ maximize social welfare $\Omega = \Sigma_n\, \Omega_n$. Hence, the LSEs' decentralized procurement decisions result in the socially optimal capacity level that does not require the ISO to know the LSEs' individual *VOLL* estimates (Woo 1990, p.77). This finding's practical importance is that while the *VOLL* estimates may be based on customer outage cost estimation studies (e.g., Woo and Pupp, 1992; Woo et al., 2014c and references thereof), they are seldom available at the hourly level for all LSEs under the ISO's jurisdiction.

2.4.2 How to buy

Absent an active market for tolling agreements, LSE *n* uses a procurement auction. The auction process has three steps (Woo et al., 2004, 2016b): (1) LSE *n* issues a request for proposal (RFP) to announce its total MW target (e.g., 1,000 MW) and eligibility criteria for IPPs' auction participation;[34] (2) interested IPPs then submit their tolling agreement offers;

[33] Appendix 1 explains LSE *n*'s calculation. There can be other demand charges for recovering LSE *n*'s transmission and distribution costs. Examples of the retail tariffs of a regulated LSE are available at https://www.pge.com/tariffs/index.page.
[34] Appendix 2 shows how LSE *n* can estimate its total MW target.

and (3) subject to cost benchmarking (Orans et al., 2004), LSE *n* selects the winning offers to execute its optimal procurement plan found in Section 2.4.1.

Aided by cost benchmarking, LSE *n*'s procurement auction is expected to yield competitively determined capacity prices (Klemperer, 2004) that track the per MW-year capacity costs of new generation units (CEC 2010). Successful Anglo-Dutch procurement auctions (Woo et al., 2003b, 2004) further allay concerns of non-competitive capacity prices in a Cournot market setting (Munoz et al., 2018). Finally, the auction may include suppliers of demand response resources (Woo et al., 2014a) to counter the potentially non-competitive behavior of IPPs (Brown, 2018b).

3. Discussion

Designing an efficient wholesale market is a well-documented challenge faced by energy economists, electrical engineers, industry practitioners, IPPs, LSEs, retail end-users, as well as regulators and policy makers. Our proposed market design meaningfully addresses the concerns of these stakeholders.

From the perspective of energy economists, a textbook model of perfect competition exemplified by an energy-only market design can have the unintended consequences of market price manipulation and inadequate incentive for generation investment. Our proposed design mitigates these consequences through its fuel-cost-based RTM price determination by an ISO and procurement auctions by LSEs under the ISO's jurisdiction.

From the perspective of electrical engineers, load-resource balance and resource adequacy are essential ingredients for reliable operation of an electric grid. Our proposed design provides such ingredients through its reliability differentiation enabled by an ISO's capacity rationing scheme and must-procure requirement for LSEs.

From the perspective of industry practitioners, a market design proposal's conceptual merit must come with practical implementation. Our proposed design is practical, relying on an ISO's existing practice of least-cost dispatch, RTM price determination and capacity rationing. Further, its suggested two-part pricing of retail consumption follows the commonly used Hopkinson tariff design (Seeto et al., 1997). Finally, it shows how to value new capacity for determining the cost-effectiveness of a demand side management program (Sreedharan et al., 2012).

From the perspective of IPPs, undesirable is a market environment that cannot provide adequate and stable revenue streams. As a result, an energy-only market's occasionally price spikes may be insufficient to induce generation investments. Our proposed design offers adequate and stable revenue streams via LSEs' procurement auctions. Further, unlike a forward contract's fuel cost risk that is borne by an IPP, a tolling agreement's fuel cost risk is borne by a LSE's retail end-users under our suggested two-part pricing of retail consumption.

From the perspective of LSEs, the goal is to best serve their retail customers. Our proposed design assumes the same goal and is therefore consistent with these LSEs' self-interests.

From the perspective of retail end-users, price reasonableness, price stability and customer choice are preferable. Our proposed design's retail pricing is reasonable based on a LSE's competitive procurement of tolling agreements and an ISO's fuel-cost-based RTM price determination. Further, the resulting RTM prices are expected to be far less volatile than those reported in California, Texas and other parts of North America (e.g., Alberta and Ontario in Canada; and New York, New England and PJM in the U.S.).

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Finally, our proposed design offers customer choice because a retail end-user can self-select its preferred reliability level through its FSL subscription and choose which LSE to be its preferred service provider.

From the perspective of regulators and policy makers, reliable service at competitive prices, customer choice, incentive compatibility, and practicality are critical criteria for gauging a market design's merit. Our proposed design meets these criteria, as demonstrated by the analysis presented in Section 2.

4. Conclusions and policy implications

We conclude by summarizing our findings reported in Section 2. First, our proposed design is a practical implementation of an ISO's least-cost dispatch, efficient RTM price determination and efficient capacity rationing. Second, it relies on market forces to achieve an electric grid's socially optimal capacity level with adequate investment incentives, albeit the ISO's ignorance of the LSEs' hourly *VOLL* estimates. Third, it does ***not*** require the ISO to operate centralized capacity auctions, use capacity payments, set high energy price caps, or subsidize market entry. Fourth, it eases the ISO's burden of market monitoring by preempting IPPs' RTM price manipulation. Finally, it suggests a LSE's two-part pricing that has a demand charge for retail FSL subscription and a real-time energy charge for retail energy consumption, thus meaningfully linking the wholesale and retail markets. When taken together, these properties have the policy implication that our proposed market design should be considered in the ongoing debate of reliability and market competition.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

We would be remiss had we failed to mention the following caveats of our proposed market design. These caveats do not invalidate our paper's key findings and policy recommendation; rather, they shape our future research agenda.

The first caveat is that some retail customers of a LSE are prosumers with behind-the-meter generation resources (e.g., roof-top photovoltaic systems) (Parag and Sovacool, 2016). Hence, the LSE's retail demand needs to be modified to include the investment behavior of a prosumer (Woo and Zarnikau, 2017).

The second caveat is that our paper does not consider the presence of a large scale development of intermittent renewable energy (e.g., solar and wind) (Coester et al., 2018; Newbery et al., 2018). Besides the approach used by Milstein and Tishler (2011), an alternative is to modify the capacity availability and merit order formulae in Section 2.2.

The third and final caveat is that we have not addressed the financial risk in an electricity grid dominated by renewable resources (Tietjen et al., 2016). A LSE's risk management, however, can be based on a profit analysis of tolling agreements (Woo et al., 2016b) and a calculation of efficient frontiers (Woo et al., 2006b). Relative to a risk-neutral LSE, a risk-averse LSE likely procures more agreements with higher capacity prices, leading to the LSE's higher expected cost but lower cost risk (Woo et al., 2004).

Acknowledgements

This paper benefits from the helpful comments of D. Brown, H.P. Chao, Y. Chen, F. Kahrl, R. Orans, and A. Yatchew. We thank M. Marchant for her copyediting and proofreading. C.K. Woo's research is supported by two grants (#4320 and #4332) from the Faculty of Liberal Arts and Social Sciences of the Education University of Hong Kong. The paper is part of J. Zarnikau's ongoing research on electricity markets at The University of Texas at Austin. Without implications, all errors are ours.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Appendix 1: LSE *n*'s retail demand charge

The calculation of LSE *n*'s retail demand charge ($/kW-month) may use one of the following three approaches:

(1) Suppose LSE *n*'s optimal procurement plan includes $K_{Jn} > 0$. Under the ISO's RTM price determination, the maximum P_h is C_{1h}. As a result, $A_J = 0$ because $\max(P_h - C_{Jh}, 0) = 0$ for all hours. LSE *n*'s marginal capacity value is $\mu_n = (\Sigma_{h \in S}\, LOLP_h \times VOLL_{hn})\, f_J = D_J$. Hence, LSE *n* sets its monthly demand charge at $\delta_n = D_J \div$ (12 months × 1000 kW per MW).

(2) Suppose $K_{Jn} = 0$ and all chosen agreements' underlying technologies have the same availability factor *f*. As $\mu_n = (\Sigma_{h \in S}\, LOLP_h \times VOLL_{hn})\, f = D_j - A_j = D - A$ for all chosen agreements, $\delta_n = (D - A) \div$ (12 months × 1000 kW per MW).

(3) Suppose $K_{Jn} = 0$ and all chosen agreements' underlying technologies have different availability factors. An estimate for μ_n is the arithmetic average of $(\Sigma_{h \in S}\, LOLP_h \times VOLL_{hn})\, f_j$, the same as the arithmetic average of $(D_j - A_j)$. Hence, δ_n = arithmetic average of $(D_j - A_j) \div$ (12 months × 1000 kW per MW).

Appendix 2: LSE *n*'s procurement target

LSE *n*'s total MW target is an estimate based on the retail customers' FSL subscriptions under the DSS's opt-in approach noted in Section 2.4.1. The estimate is (MW_{FSL} / f), where MW_{FSL} = total MW size of all FSL subscriptions unmet by LSE *n*'s resources already in place; and f = average availability factor of the generation technologies likely used by the interested IPPs in Step 2.

Since CTs and CCGTs dominate new plant construction in North America, an estimate for *f* is 0.95 based on these generation units' forced outage rates reported in CEC

(2010, p.C-10). This high f estimate implies that a region's resource adequacy requirement (RAR) can be as low as $(1 / f) = (1 / 0.95) \approx 106\%$, well below California's current RAR of 115% (Woo et al., 2016b, p.52).

A large decrease in a region's RAR, however, is likely imprudent because the target estimation portrayed here has not considered transmission-related issues such as line failures, congestion and network security. Nevertheless, a potentially large RAR reduction highlights the resource benefit of knowing, in advance and with certainty, a retail customer's kW load (= FSL subscription) in a shortage hour.

References

Alagappan, L., Orans, R., Woo, C.K., 2011. What drives renewable energy development? Energy Policy 39, 5099-5104.

Bohn, R.E., Caramanis, M.C., Schweppe, F.C., 1984. Optimal pricing in electrical networks over space and time. Rand Journal of Economics 15, 360-376.

Borenstein, S., Bushnell, J.B., Wolak, F.A., 2002. Measuring market inefficiencies in California's restructured wholesale electricity market. American Economic Review 92(5), 1376-1405.

Brown, D.P., Olmstead, D.E.H., 2017. Measuring market power and the efficiency of Alberta's restructured electricity market: an energy-only market design. Canadian Journal of Economics 50(3), 838-870.

Brown, D.P., 2018a. The effect of subsidized entry on capacity auctions and the long-run resource adequacy of electricity markets. Energy Economics 70, 205-232.

Brown, D.P., 2018b. Capacity payment mechanisms and investment incentives in restructured electricity markets. Energy Economics 74, 131-142.

Bublitz, A., Keles, D., Zimmermann, F., Fraunholz, C., Fichtner, W., 2019. A survey on electricity market design: insights from theory and real-world implementation of capacity remuneration mechanisms. Energy Economics, in press.

CEC, 2010. Comparative Costs of California Central Station Electricity Generation. Report CEC- 200-2009-07SF, California Energy Commission, Sacramento, California. Available at: http://www.energy.ca.gov/2009publications/CEC-200-2009-017/CEC-200-2009-017-SF.PDF (Accessed on 6 March 2019).

Chao, H.P., 1983. Peak load pricing and capacity planning with demand and supply uncertainty. Bell Journal of Economics 14(1), 179-190.

Chao, H.P., Wilson, R., 1987. Priority service: pricing, investment and organization. American Economic Review 77(4), 899-916.

Chao, H.P., Wilson, R., 2004. Resource adequacy and market power mitigation via option contracts. Available at: https://faculty-gsb.stanford.edu/wilson/PDF/Electricity%20Markets/Chao-Wilson,Resource%20Adequacy%20and%20Market%20Power%20Mitigation031804.pdf (Accessed on 27 February 2019).

Chao, H.P., 2012. Competitive electricity markets with consumer subscription service in a smart grid. Journal of Regulatory Economics 41(1), 155-180.

Coester, A., Hofkes, M.J., Papyrakis, E., 2018. An optimal mix of conventional power systems in the presence of renewable energy: a new design for the German electricity market. Energy Policy 116, 312-322.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Conejo, A.J., Sioshansi, R., 2018. Rethinking restructured electricity market design: lessons learned and future needs. International Electrical Power & Energy Systems 98, 520-530.

Cramton, P., 2017. Electricity market design. Oxford Review of Economic Policy 33(2), 589-612.

Deng, S.J., Johnson, B., Sogomonian, A., 2001. Exotic electricity options and valuation of electricity generation and transmission. Decision Support Systems 30(3), 383–392.

Deng, S.J., Oren, S.S., 2006. Electricity derivatives and risk management. Energy 31(6–7), 940–953.

Distributed Energy Financial Group, 2015. The Annual Baseline Assessment of Choice in Canada and the United States (ABACCUS). Available at: http://defgllc.com/publication/abaccus-2015-annual-baseline-assessment-of-choice-in-canada-and-the-united-states/ (accessed on July 20, 2018).

Eydeland, A., Wolyniec, K., 2003. Energy and Power Risk Management: New Development in Modeling, Pricing and Hedging. John Wiley & Sons, Hoboken, New Jersey.

Hobbs, B.F., 1995. Optimization methods for utility resource planning. European Journal of Operational Research 83(1) 1-20.

Faruqui, A., Sergici, S., 2010. Household response to dynamic pricing of electricity: a survey of 15 experiments. Journal of Regulatory Economics 38(2), 193–225.

Gal, N., Milstein, I., Tishler, A., Woo, C.K., 2017. Fuel cost uncertainty, capacity investment and price in a competitive electricity market. Energy Economics 61, 233-240.

Gal, N., Milstein, I., Tishler, A., Woo, C.K., 2019. Investment in electricity capacity under fuel cost uncertainty: dual-fuel and a mix of single-fuel technologies. Energy Policy 126, 518-532.

Joskow, P.L., Tirole, J., 2007. Reliability and competitive electricity markets. Rand Journal of Economics 38(1), 60-84.

Joskow, P.L., 2013. Editorial: symposium on capacity markets. Economics of Energy & Environmental Policy 2(2), v-vi.

Klemperer, P., 2004. Auctions: Theory and Practice. Princeton University Press, Princeton, New Jersey.

Kumar, A., Srivastava, S.C., Singh, S.N., 2005. Congestion management in competitive power market: a bibliographical survey. Electric Power Systems Research 76, 153-164.

Laffont, J.J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, Massachusetts.

Lusztig, C., Feldberg, P., Orans, R., Olson, A., 2006. A survey of transmission tariffs in North America. Energy 31, 1017-1039.

Milstein, I., Tishler, A., 2011. Intermittently renewable energy, optimal capacity mix and prices in a deregulated electricity market. Energy Policy 39, 3922-3927.

Milstein, I., Tishler, A., 2012. The inevitability of capacity underinvestment in competitive electricity markets. Energy Economics 34(1), 62-77.

Milstein, I., Tishler, A., 2019. On the effects of capacity payments in competitive electricity markets: capacity adequacy, price cap, and reliability. Energy Policy 129, 370-385.

Moore, J., Woo, C.K., Horii, B., Price, S., Olson, A., 2010. Estimating the option value of a non-firm electricity tariff. Energy 35, 1609-1614.

Munoz, F.D., Wogrin, S., Oren, S.S., Hobbs, B.F., 2018. Economic inefficiencies of cost-based electricity market designs. The Energy Journal 39(3): 51 – 68.

Newbery, D., Pollitt, M., Ritz, R., Strielkowski, W., 2018. Market design for a high-renewables European electricity system. Renewable and Sustainable Energy Reviews 91, 695-707.

Newsham, G.R., Bowker, B.G., 2010. The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: a review. Energy Policy 38(7), 3289 – 3296.

Orans, R., Woo, C.K., Clayton, W., 2004. Benchmarking the price reasonableness of a long-term electricity contract. Energy Law Journal 25(2), 357-383.

Parag, Y., Sovacool, B., 2016. Electricity market design for the prosumer era. Nature Energy 1, article number: 16032.

Seeto, D.Q., Woo, C. K., Horowitz, I., 1997. Time-of-use rates vs. Hopkinson tariffs *redux*: an analysis of the choice of rate structures in a regulated electricity distribution company. Energy Economics 19, 169-185.

Sioshansi, F.P., 2013. Evolution of Global Electricity Markets: New Paradigms, New Challenges, New Approaches. Elsevier, San Diego, California.

Spees, K., Newell, S.A., Pfeifenberger, J.P., 2013. Capacity markets – lessons learned from the first decade. Economics of Energy & Environmental Policy 2(2), 1-26.

Sreedharan, P., Miller, D., Price, S., Woo, C.K., 2012. Avoided cost estimation and cost-effectiveness of permanent load shifting in California. Applied Energy 96, 115-121.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Stern, L., 1998. Merchant power plants: competing for financing in a deregulated industry. The Journal of Structured Finance 4(3), 47–55.

Stoft, S., 2002. Power System Economics: Designing Markets for Electricity. Wiley-IEEE Press, New York, New York.

Tietjen, O., Pahle, M., Fuss, S., 2016. Investment risks in power generation: a comparison of fossil fuel and renewable energy dominated markets. Energy Economics 58, 174-185.

Wolfram, C.D., 1999. Measuring duopoly power in the British electricity spot market. American Economic Review 89(4), 805-826.

Woo, C.K., 1990. Efficient electricity pricing with self-rationing. Journal of Regulatory Economics 2(1), 69-81.

Woo, C.K., 1991. Capacity rationing and fixed cost collection. The Energy Journal 12(2), 153-164.

Woo, C.K., Pupp, R.L., 1992. Costs of service disruptions to electricity consumers. Energy 17(2), 109-126.

Woo, C.K., 1993. Efficient electricity pricing with self-rationing: reply. Journal of Regulatory Economics 5(1), 101-102.

Woo, C. K., Horowitz, I., Martin J., 1998. Reliability differentiation of electricity transmission. Journal of Regulatory Economics 13, 277-292.

Woo, C.K., Lloyd, D., Tishler, A., 2003a. Electricity market reform failures: UK, Norway, Alberta and California. Energy Policy 31(11), 1103-1115.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Woo, C.K., Borden, M., Warrington, R., Cheng, W., 2003b. Avoiding overpriced risk management: exploring the cyber auction alternative. Public Utilities Fortnightly 141(2), 30-37.

Woo, C.K., Karimov, R., Horowitz, I., 2004. Managing electricity procurement cost and risk by a local distribution company. Energy Policy 32(5), 635-645.

Woo, C.K., King, M., Tishler, A., Chow, L.C.H., 2006a. Costs of electricity deregulation. Energy 31(6-7), 747-768.

Woo, C.K., Horowitz, I., Olson, A., Horii, B., Baskette, C., 2006b. Efficient frontiers for electricity procurement by an LDC with multiple purchase options. OMEGA 34(1), 70-80.

Woo, C.K., Lloyd, D., Clayton, W., 2006c. Did a local distribution company procure prudently during the California electricity crisis? Energy Policy 34(16), 2552-2565.

Woo, C.K., Kollman, E., Orans, R., Price, S., Horii, B., 2008. Now that California has AMI, what can the state do with it? Energy Policy 36, 1366-1374.

Woo, C.K., Horowitz, I., Moore, J., Pacheco, A., 2011. The impact of wind generation on the electricity spot-market price level and variance: the Texas experience. Energy Policy 39, 3939-3944.

Woo, C.K., Zarnikau, J., Kadish, J., Horowitz, I., Wang, J., Olson, A., 2013. The impact of wind generation on wholesale electricity prices in the hydro-rich Pacific Northwest. IEEE Transactions on Power Systems 28(4), 4245-4253.

Woo, C.K., Sreedharan, P., Hargreaves, J., Kahrl, F., Wang, J., Horowitz, I., 2014a. A review of electricity product differentiation. Applied Energy 114, 262-272.

Woo, C.K., Ho, T., Zarnikau, J., Olson, A., Jones, R., Chait, M., Horowitz, I., Wang, J., 2014b. Electricity-market price and nuclear power plant shutdown: evidence from California. Energy Policy 73, 234-244.

Woo, C.K., Ho, T., Shiu, A., Cheng, Y.S., Horowitz, I., Wang, J., 2014c. Residential outage cost estimation: Hong Kong. Energy Policy 72, 204-210.

Woo, C.K., Moore, J., Schneiderman, B., Ho,T., Olson, A., Alagappan, L., Chawla, K., Toyama, N., Zarnikau, J., 2016a. Merit-order effects of renewable energy and price divergence in California's day-ahead and real-time electricity markets. Energy Policy 92, 299-312.

Woo, C.K., Horowitz, I., Zarnikau, J., Moore, J., Schneiderman, B., Ho, T., Leung, E., 2016b. What moves the *ex post* variable profit of natural-gas-fired generation in California? The Energy Journal 37(3), 29-57.

Woo, C.K., Chen, Y., Olson, A., Moore, J., Schlag, N., Ong, A., Ho, T., 2017a. Electricity price behavior and carbon trading: new evidence from California. Applied Energy 204, 531-543.

Woo, C.K., Olson, A., Chen, Y., Moore, J., Schlag, N., Ong, A., Ho, T., 2017b. Does California’s CO_2 price affect wholesale electricity prices in the western U.S.A.? Energy Policy 110, 9-19.

Woo, C.K., Zarnikau, J., Shiu, A., Li, R., 2017c. Winter residential optional dynamic pricing: British Columbia, Canada. The Energy Journal 38(5), 99-111.

Woo, C.K., Zarnikau, J., 2017. A solar rate option for the development of behind-the-meter photovoltaic systems. The Electricity Journal 30(3), 1-3.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

Woo, C.K., Chen, Y., Zarnikau, J., Olson, A., Moore, J., Ho. T., 2018a. Carbon trading's impact on California's real-time electricity market prices. Energy 159, 579-587.

Woo, C.K., Zarnikau, J., Chen, Y., Olson, A., Moore, J., Ho, T., Liu, Y., Luo, X., 2018b. An empirical analysis of California's hybrid capacity options. The Electricity Journal 31(2), 7-12.

Woo, C.K., Liu, Y., Zarnikau, J., Shiu, A., Luo, X., Kahrl F., 2018c. Price elasticities of retail energy demands in the United States: new evidence from a panel of monthly data for 2001 – 2016. Applied Energy 222, 460-474.

Wood, A.J., Wollenberg, B.F., 2012. Power Generation, Operation, and Control. Wilely, New York, New York.

Zarnikau, J., 2011. Successful renewable energy development in a competitive electricity market: a Texas case study. Energy Policy 39(7), 3906–3913.

Zarnikau, J., Woo, C.K., Baldick, R., 2014. Did the introduction of a nodal market structure impact wholesale electricity prices in the Texas (ERCOT) market? Journal of Regulatory Economics 45, 194-208.

Zarnikau, J., Woo, C.K., Zhu, S., Tsai, C.H., 2019. Market price behavior of wholesale electricity products: Texas. Energy Policy 125, 418-428.

This is the pre-published version published in Energy Policy, available online at: https://doi.org/10.1016/j.enpol.2019.110988.

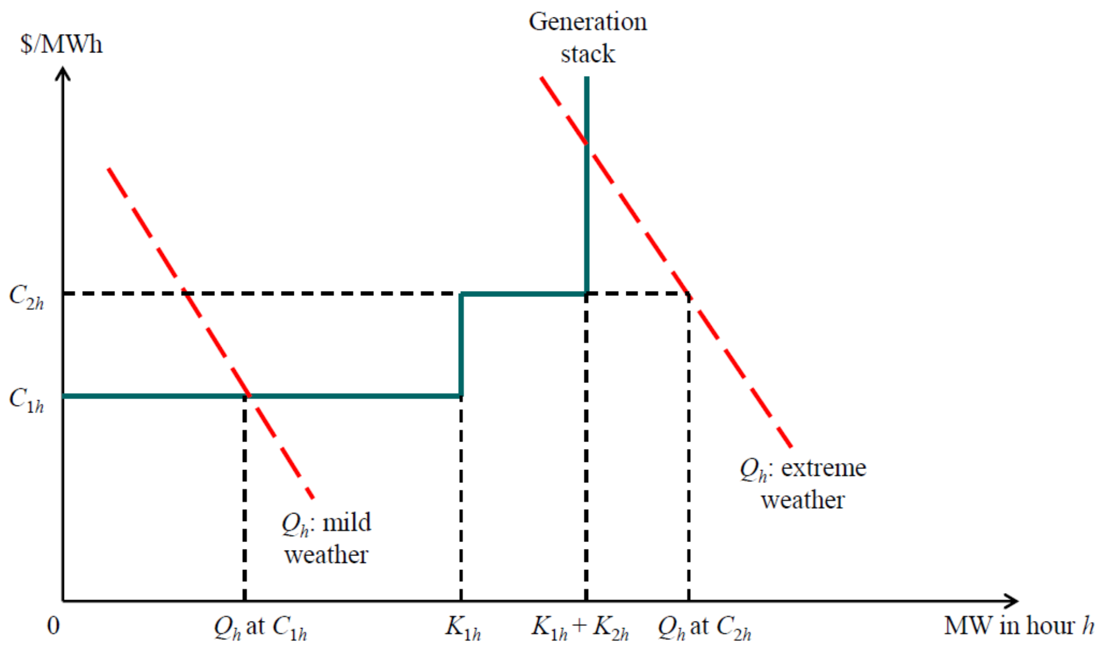

Figure 1. The ISO’s least-cost dispatch, RTM price determination and capacity rationing