Electricity Market Design

Peter Cramton[1]

25 July 2017

Abstract

Electricity markets are designed to provide reliable electricity at least cost to consumers. This paper describes how the best designs satisfy the twin goals of short-run efficiency—making the best use of existing resources—and long-run efficiency—promoting efficient investment in new resources. The core elements are a day-ahead market for optimal scheduling of resources and a real-time market for security-constrained economic dispatch. Resources directly offer to produce per their underlying economics and then the system operator centrally optimizes all resources to maximize social welfare. Locational marginal prices, reflecting the marginal value of energy at each time and location, are used in settlement. This spot market provides the basis for forward contracting, which enables participants to manage risk and improves bidding incentives in the spot market. There are important differences in electricity markets around the world, reflecting different economic and political settings. Electricity markets are undergoing a transformation as the resource mix transitions from fossil fuels to renewables. The main renewables, wind and solar, are intermittent, have zero-marginal cost, and lack inertia. These challenges can be met with battery storage and improved demand response. However, good governance is needed to assure the market rules adapt to meet new challenges.

JEL: D44, D47, L94

Keywords: market design, electricity markets, electricity regulation, smart markets

[1] University of Maryland, European University Institute, and University of Cologne, e-mail: pcramton@gmail.com
Since 1983, Peter Cramton has conducted research on market design; he has applied that research to design auction-based markets of radio spectrum, electricity, financial securities, and other products. I thank the editors and referees for helpful comments, as well as my long-term collaborators on this topic, Axel Ockenfels, Steven Stoft, and Robert Wilson. Financial support by the University of Cologne through the Key Profile Area II is gratefully acknowledged.

Introduction

Today's restructured electricity markets illustrate the importance and power of effective market design. Over the last twenty-five years, electricity markets have evolved to address complex economic and engineering challenges. Despite some bumps along the way, the markets have largely succeeded in the goal of providing reliable electricity at least cost to consumers. This is no simple task. Every second supply and demand must balance. Thousands of resource and network constraints must be satisfied. And the market must send the right price signals to motivate efficient generation and investment in resources over time.

The complexity of the economic problem that the market must solve necessarily makes the market design complex. My purpose here is to present the basic features of the electricity market, explain how the market works to address key objectives, how it has evolved over time, and whether it is suited to handle the large changes that will take place over the next decades.

Not surprisingly, electricity markets differ in their design in different regions. Some differences reflect key differences in the market setting, but differences also arise for path-dependent reasons. Wilson's (2002) Presidential Address to the Econometric Society studies two alternative designs: (1) an integrated market in which the system operator centrally optimizes the scheduling and dispatch of resources, and (2) an exchange-based market in which energy companies trade day-ahead and throughout the day at prices that clear the market absent central optimization. I focus primarily on the integrated market, as it has become the standard market design in North America (FERC 2002) and elsewhere. However, I do describe key differences with markets in Europe, which tend to follow the exchange-based market model. I also point out key differences within the US markets, such as the inclusion of an explicit capacity market in addition to the energy market.

Electricity markets are designed markets. They did not emerge from an unorganized marketplace. Rather they were designed in a regulatory process in part because electricity is viewed as an essential service, but also because of its technical properties. Electricity began as a monopoly utility, then shifted to a power pool in which the monopoly utilities could engage in trade. The final step to markets came with the introduction of spot markets that determine the quantities generated and consumed as well as the prices paid for energy and related services at each time and location.

Good electricity market design has always been important. Design mistakes can cost consumers tens of billions of dollars, as illustrated by the California electricity crisis of 2000-2001 (Borenstein 2002). Fortunately, because of good governance and technological progress, market designs have improved over time. Flaws have been identified and largely addressed.

Still electricity market design is far from static. New challenges are emerging with the ongoing transformation of the electricity industry. The forces driving change are the expansion of renewables, demand response, distributed generation, smart homes, and battery storage.

Electricity is front and center in debates on climate change. Electricity generation is the single largest emitter of carbon to the atmosphere. Efforts to address climate change will shift generation from fossil fuels to wind, solar, nuclear and other non-emitting resources. The electricity market design must be able to handle this transformation. The task is non-trivial as the main renewable resources, wind and solar, are intermittent sources of supply with zero marginal cost of production and no inertia. Today's markets can easily handle a moderate share of renewable generation, but what if generation is dominated by renewables? Are adjustments to the market design needed to handle such major shifts in the generation mix?

Given the strong foundation laid over the last two-decades, I am optimistic that electricity market designs will continue to improve and meet the coming challenges. The key is good governance, forward planning, and a strong focus on basic market principles to achieve market objectives.

Market objectives

In broadest terms, regulators seek a market design that provides reliable electricity at least cost to consumers. This can be broken down into two key objectives.

The first is short-run efficiency: making the best use of existing resources. Although complex due to the number of resources, the economics of the resources, and the constraints on the system, this objective is achieved through direct optimization. Provided the economics of resources are truthfully bid and absent other distortions, then the centrally optimized unit commitment and real-time dispatch will achieve an efficient welfare-maximizing outcome.

This ideal is not achieved in practice. Efforts to mitigate market power are imperfect. Emission externalities are improperly priced. And a myriad of subsidies and command and control regulations are imposed. Given these distortions, one might argue that the optimization of resources is naïve—it will invariably come up with the wrong answer. I, however, believe that the transparent and direct optimization of resources, helps identify distortions, such as market power, so that they can be addressed. If carbon emissions are underpriced, then the solution is to properly price them, rather than to alter the market design to disadvantage coal generation in some non-transparent way.

The second objective is long-run efficiency: ensuring the market provides the proper incentives for efficient long-run investment. This has proven to be the most challenging objective. In the simplest theory, efficient long-run investment is induced from the right spot prices. But this is complicated by the reliability requirement.

Reliability requires a reserve to satisfy demand when supply and demand uncertainty would otherwise lead to shortage. In other industries, reliability is not an issue. Prices rise and fall to assure supply and demand balance, but in current electricity markets there typically is insufficient demand that responds to price and consumers are unable to express a preference for reliability. Thus, there is a need in current markets for the regulator to determine how this preference for reliability is expressed. As we will see, one approach to reliability is to rely solely on spot prices but to include administrative scarcity prices at times when reserves are scarce. The preference for reliability is imbedded in the scarcity prices. Setting higher scarcity prices enhances reliability in providing stronger investment incentives. An alternative approach is to more directly coordinate investment with a capacity market, although this is best done as an addition to, not a substitute for, administrative scarcity pricing, since it is the scarcity price that motivates capacity to perform when needed.

Long-run efficiency remains the most challenging and important of market objectives. The drive to restructure markets emerged from poor investment decisions under rate-of-return regulation. Restructured markets provide strong incentives for sound investment, since it is the investor who bears the financial consequence of the decisions, not the rate payer. In competitive markets, consumers ultimately benefit from these better investment decisions.

There are other market objectives, such as simplicity, transparency, and fairness.

Electricity markets are necessarily complex (see Stoft 2002 for a good discussion). This follows from the complexity of the engineering and economic problem that must be solved. Still designers should strive to keep the design as simple as possible. Complicating features should only be added if they are necessary and consistent with market principles.

Electricity markets have a high degree of transparency. Market rules and their development and review are publicly available. Market data is available in real time and periodically reviewed month, quarterly, and annually. The planning process also has a high level of transparency. Transparency helps identify and address problems. It also supports efficient operation and investment.

A key element of fairness is equal treatment and open access to the market. Fairness is encouraged with the independence of the system operator and a governance structure that includes representation of all stakeholders.

Market evolution

Electricity began as a monopoly utility. Each region had its own electric utility that would provide generation of electricity, transmission on high-voltage lines, and finally distribution on low-voltage lines to homes and businesses. Under rate-of-return regulation, the utility would recommend investments to regulators and if approved, the costs would be included in the rate base. The entire operation was run centrally by a single entity.

The first step toward markets was the creation of power pools. In a power pool, several neighboring utilities are connected via the transmission network, allowing the trade of energy across regions. Trade has both cost and reliability benefits. First, drawing from a larger fleet of generators means that the required energy can be supplied at lower cost, as low-cost resources are used more and high-cost resources are used less. Second, access to the capacity in other regions makes it easier for a region to supply energy when a resource fails or demand spikes. Thus, the system becomes more reliable and fewer reserves are needed. However, in early power pools, the gains from this interconnection were limited because of the absence of a robust spot market.

The final step to markets is the creation of a wholesale market. The wholesale market allows real-time trade and pricing of energy.

Most markets began as single-price markets. In such markets transmission constraints are ignored for purposes of pricing, but of course cannot be ignored in physical delivery. To satisfy constraints, some out-of-merit generation is asked to produce, displacing some in-merit generation. This approach only works well if there is sufficient transmission so that constraints rarely bind. Otherwise, the approach has poor incentives. To address these incentive problems, nodal markets often are introduced. With nodal markets, energy prices properly reflect transmission constraints. The price equals the marginal value of energy at each time and location.

The shift to nodal markets is just one of many enhancements to current markets. Market rules are constantly being improved to address observed flaws and to respond to new challenges. Good governance is essential for the markets to see steady improvement.

Market structure

The cornerstone of a restructured electricity market is the wholesale market in which generators compete to serve load (the demand side of the market). Generation is effectively unbundled from the electric utility. This allows generators to operate in a competitive market in which they make their own investment decisions and stand to gain or lose based on how those decisions turn out.

Retail competition is a second element of the market model. Here the service providers, commonly called load serving entities, compete for retail customers. One might think competition would be intense, because price is the salient attribute of the retail service. However, even though customers have a strong preference to minimize cost, most retail customers are poor electricity shoppers, requiring education and objective information to identify the best contract (Wilson and Price 2010, Competition and Markets Authority 2016).

As we look ahead, service providers will also compete on plans and technologies that help customers minimize electricity cost subject to their preferences. The coming smart home will enable both the shifting and reduction of demand for the benefit of customers. Innovative service providers that do the best job of maximizing customer value will prosper. Retail competition will become increasingly important as smart home technologies are developed and adopted.

Transmission is a critical yet difficult element in the market. Transmission is critical in that it is transmission that enables the wholesale market. Indeed, an important benefit of transmission is that it enhances competition in the wholesale market, thereby allowing the market to work better. However, transmission investment is not a problem that the market solves (Chao, et al. 2008). One difficulty is that forward-looking congestion rents are an inadequate means of cost recovery for lumpy transmission investments. Efficient new transmission lines often eliminate the congestion rents that would otherwise motivate the investment. Cost recovery is possible through additional charges, for example with Ramsey pricing, but transmission planning and investment occurs as part of the regulatory process.

An unresolved problem is managing the interaction between transmission investment and generation investment. Optimal generation investments require knowledge of long-run transmission plans, but the generation investments ultimately influence the transmission plans. There is no obvious answer to this coordination problem.

The final element in the market model is distribution, the low-voltage lines that bring electricity to our homes and businesses. The distribution company remains a monopoly utility in the restructured market.

But even here the landscape is changing as various forms of distributed generation and storage are introduced.

A successful market design

There are many different electricity market designs. Here I outline the standard market design used in most of North America, which has proven successful (O'Connor et al. 2015). Even within North America, there are important differences among the markets. For concreteness, I focus on the Texas market, run by its independent system operator, the Electric Reliability Council of Texas (ERCOT).[2]

ERCOT is a large and growing market with a peak demand of 71 GW (summer 2016), covering 24 million people or 90% of the electricity demand in Texas. In 2016, 99.5% of the energy came from four generation-types: 43.7% natural gas, 28.8% coal, 15.1% wind, and 12% nuclear (ERCOT 2017). New build is dominated by wind and solar. ERCOT has limited interconnection via five DC ties to other markets in the US (820 MW) and Mexico (430 MW). Other characteristics of ERCOT are excellent wind and solar potential, large industrial demand, and a responsive planning process for both generation and transmission. Due to low natural gas prices and ample capacity (78 GW), the average wholesale energy price reached an all-time low of $24.62/MWh for 2016.

The market design must address several challenges. First, both the transmission grid and each generating resource has physical limits. Transmission lines can carry only so much power, and power flows per physical laws (Kirchhoff's Laws). Generating resources have numerous limits in how they can be operated. These constraints must be respected to prevent the physical destruction of equipment. Second, there is significant uncertainty in both supply and demand. Generators and network elements fail at times. Demand varies in both predictable and unpredictable ways, as does intermittent supply, such as wind and solar generation. The market must accommodate this uncertainty. Third, supply and demand must be balanced at every moment, yet much of demand does not respond to price, making it harder to clear the market.

Although the market is centralized, it relies on the decentralized decisions of market participants. Central market clearing and dispatch allow the system operator to meet the twin objectives of reliability and efficiency.

The day-ahead and real-time optimization of resources involves state-of-art optimization techniques and hardware. To get a sense of the magnitude of the problem, ERCOT has thousands of computer servers to run its systems. This is very much a smart market. Preferences and constraints are expressed in sophisticated ways and then optimized to achieve the highest welfare possible. The case for the integrated market model has strengthened over the last twenty years. There has been enormous technological progress, allowing much more realistic and powerful optimization of resources.

The two core market elements are the day-ahead market and the real-time market, which I collectively call the spot market. In the day-ahead market, participants submit bids and offers for energy and reserves

[2] Disclosure: I am an independent director on the ERCOT Board. The views expressed are my own and do not necessarily reflect those of the ERCOT Board.

for each hour of the next day. The outcome is a day-ahead schedule of supply to meet demand together with prices for each hour and location. The outcome is financially, but not physically, binding. The day-ahead market lets participants efficiently coordinate production as they plan for the next day and hedge more volatile real-time prices. The real-time market is a bid-based, security constrained economic dispatch that is conducted at least every five minutes throughout the day. The outcome is the efficient physical dispatch of resources together with 5-minute prices at each location. (See Hogan 2014 for a discussion of some of the pricing challenges.) I now discuss these core market elements in greater detail.

Day-ahead market

The day-ahead market is voluntary and financially binding. Participants submit bids and offers to buy and sell energy and ancillary services (reserves of various forms). Each day at 0600 the system operator publishes system conditions, forecasts, and other information for the next day. Participants submit bids and offers by 1000. The day-ahead outcome is announced by 1330. The results help operators and market participants plan for the next day and provide incentives for efficient scheduling of units (unit commitment) and operation throughout the day.

An offer of a generator is expressed in three parts: startup cost, minimum-energy cost, and an energy offer curve. The startup cost reflects the cost to start the generator. The minimum-energy cost is the cost to run the generator at its lowest sustainable level. The energy offer curve is a piecewise-linear increasing offer curve defined by up to ten price-quantity pairs—it reflects the unit's marginal cost for energy beyond the minimum level.

An important advantage of these three-part offers is they let the resource express its underlying economics. This enables the efficient commitment of resources in competitive situations. In uncompetitive situations, rules mitigate the exercise of market power. Some early designs only allowed an energy offer curve. A generator then would have to internalize startup and minimum-energy costs in its offer based on expectations of how the resource might be scheduled. This guesswork often would be wrong and would distort the energy offer curve. To avoid gaming of the startup and minimum-energy offers, these components are capped at generic or verified costs. Other parameters of the resource such as ramp rates and low and high limits reflect the resource's physical capability.

The day-ahead market also allows virtual offers and bids that are not tied to a generating unit. A key purpose of these offers and bids is to enable arbitrage between the day-ahead and real-time markets. A trader expecting higher prices in the real-time market can buy day-ahead and then sell in real time. If the trader is right, the trader makes money and causes the day-ahead and real-time prices to converge.

Virtual bids also allow generators to offer a resource's true economics in the day-ahead market without limiting the generator's ability to sell the energy where it expects to get the best price. A generator expecting a higher real-time price can offer the resource at cost in the day-ahead market to get efficiently scheduled day-ahead, and buy day-ahead with a virtual bid to shift the sale to the real-time market. Empirical studies confirm the benefits of virtual bidding (Jha and Wolak 2014).

The day-ahead market solves a large optimization to determine the quantities traded and the prices. The objective is to maximize the as-bid gains from trade (total value minus total cost) subject to the network and resource constraints. In the simplest case, each supply resource has an increasing marginal cost curve, each demander has a decreasing marginal value curve, and all constraints on resources are linear. The optimization then is a standard convex optimization. The outcome is a competitive equilibrium with supporting prices computed from the shadow prices on constraints. Each supplier sells its profit maximizing quantity at each time and location given the prices and the physical constraints and each demander buys its profit maximizing quantity at each time and location, given the prices and constraints.

The energy prices are locational marginal prices (LMPs) representing the additional cost of generating one additional megawatt at the location in the hour. These linear prices are used for settlement. Sellers are paid price times quantity and buyers are charged price times quantity. Price differences arise because of binding transmission constraints. The price difference between any two locations (price at sink minus price at source) is the congestion rent. Congestion revenue between two points is the price difference times quantity.

The actual day-ahead optimization is not so simple. There are many non-convexities that must be addressed. The startup and minimum-energy costs are examples of non-convexities. As a result, the optimization is a mixed-integer programming problem. Fortunately, the hardware and software for solving such problems continues to improve, allowing the solution of more complex models with fewer simplifying assumptions. Often the problems are nearly convex at least around the optimum. Still a competitive equilibrium with supporting prices may not exist. The optimizer instead finds the optimal quantities and approximate prices. Ancillary services are co-optimized with energy, so the solution finds optimal quantities for energy and ancillary services. Since the prices are not true supporting prices, it is possible that over the day a generator selected to run does not earn enough energy and ancillary service revenue to cover its total as-bid costs including startup and minimum-energy costs. In this case, the generator receives a make-whole payment to cover any shortfall. (This may alter bidding incentives, but recall gaming of the startup and minimum energy offers is mitigated with a requirement to offer generic or verified costs.) The make-whole payments are charged to load that clears in the day-ahead market. In practice, make-whole payments are an insignificant portion of costs, suggesting that the non-convexities, while important to address, are not too disruptive.

Following the release of the day-ahead results (1430), participants report any bilateral trades of energy, capacity, or ancillary services for the next day. These trades impact the settlement.

Also by 1430, each participant updates and submits its current operating plan for the next seven days. The operating plan is mandatory. It reports anticipated operating conditions of the participant’s resources. Participants update the plans within 60 minutes of changes. Modifications are allowed until an hour ahead of the operating hour. Transmission service providers also submit plans reporting any planned or forced transmission outages.

The system operator uses the current operating plans to assure sufficient resources are available to reliably meet demand at each time and location. By 1800 day-ahead, the system operator analyzes transmission security and runs a reliability unit commitment to determine if it is necessary to commit any additional units to avoid a security violation—the violation of a constraint in response to a contingency (the failure of a resource or transmission line). So as not to interfere with the market, the system operator waits as long as possible before making reliability unit commitments. These commitments are rare in practice. Further, these units have an offer floor of $1500/MWh for energy above the low sustainable limit, so they only will be used for energy when they are absolutely needed. The intent is to assure reliability with minimal interference in the market.

Adjustment period

From 1800 day-ahead until 60 minutes before the operating hour, there is an adjustment period in which participants report any changes to their current operating plans. The system operator uses the updated information to rerun the transmission security analysis and the reliability unit commitment to determine if any additional units are needed to avoid security violations.

Throughout the adjustment period, the system operator also procures any additional ancillary services that are required because of changes in plans for the operating day.

Real-time market

The real-time market determines the dispatch of resources and pricing of energy throughout the operating day. The real-time market is mandatory. Resources are required to have an energy offer curve or, for self-scheduled units, an output schedule. An output schedule specifies the resource's output in every five-minute interval. The self-scheduled resource is a price taker for energy at the scheduled quantity.

Load resources that can respond to dispatch instructions can submit energy bids. The bid curves take the same form as the energy offer curves, but are decreasing—less is demanded at higher prices. This demand response is playing an increasingly important role in the balancing of supply and demand. It is especially effective in Texas, which has substantial industrial load that is well-suited to respond to prices in times of scarcity. Demand response also comes from load aggregators.

Every five minutes, the system operator takes the current system state and energy offer and bid curves and runs the security-constrained economic dispatch. This results in real-time dispatch instructions for all resources together with prices. The real-time dispatch is the least-cost dispatch of all online resources that satisfies transmission and generation constraints.

In real-time, market power becomes a more severe problem as the system operator has fewer options—resources are limited to those online and the ability of the resources to react is limited by ramp rates. Some method of mitigating market power is required. ERCOT's procedure is to run the security-constrained economic dispatch (SCED) twice, first with a partial set of constraints that are deemed competitive, and then with the full set of constraints. The LMPs generated in the first (competitive) run are used as reference prices. In the second run with all constraints, the energy offer curve of each online resource is capped at the higher of the reference LMP or the mitigated offer cap. The mitigated offer cap is computed from a fuel price index and a generic heat rate to approximate the resource's marginal cost.

A constraint is deemed competitive if 1) the market is not too concentrated on the import side of the constraint, 2) the market is not too concentrated on the export side of the constraint, and 3) no supplier is pivotal (can cause violation of the constraint unilaterally). Market concentration is measured with the

Herfindahl-Hirschmann index. A supplier's market share is weighted by the shift factor, which measures the supplier's impact on the constraint. A supplier must have a shift factor exceeding 2% to be included in the calculation. The list of non-competitive constraints is determined and reported daily.

When a transmission constraint binds, the optimizer buys counterflow to satisfy the constraint. Absent mitigation, a supplier located to sell counterflow on a non-competitive constraint could inflate offers to secure a high price. The mitigation prevents these inflated offers. Notice that the dispatch and pricing only change when non-competitive constraints bind.

All North American markets have some form of market power mitigation in the real-time market. For example, PJM uses the three-pivotal supplier test: if the combined resources of the three largest suppliers at a constraint are pivotal, then their offers are mitigated to cost-based offers—offers set administratively to reflect cost plus 10%. PJM applies this rule in all its markets to address local market power, not just the real-time market. Texas is unusual in that it does not mitigate offers in the day-ahead market.

The real-time market is financially and physically binding. Resources must follow dispatch instructions. Failure to follow within a tolerance results in a non-compliance penalty. The tolerance for intermittent resources (wind and solar) is +10%. For other resources, it is the greater of ±5% or ±5MW.

If supply and demand were certain, the real-time market would be unnecessary. Participants could make optimal plans in the day-ahead market and then stick to them. The beauty of the real-time market is that it allows participants to make efficient adjustments as uncertainty is resolved. The real-time market is an efficient means of settling deviations from the day-ahead plans.

To clarify the interaction between the day-ahead and real-time markets, consider the following typical example. A resource without market power submits a three-part offer day-ahead to be efficiently scheduled throughout the day. The day-ahead outcome is financially binding. Nothing changes during the day for the resource, so it maintains its offer throughout the operating day and is dispatched in real time based on the offered economics. The motivation for offering this flexibility in the real-time market is to capture additional gains from trade. The resource may be asked to generate more when price exceeds marginal cost and less when price is below marginal cost, resulting in additional profits.

Ancillary services

Energy markets require reserves of various forms, collectively called ancillary services, to balance supply and demand in every second and satisfy all constraints. In Texas, there are three types: regulation, responsive reserve, and non-spinning reserve. (Some markets have many more ancillary services. The need for certain services may stem from market design deficiencies or features of the market, such as a high level of intermittent resources.) Regulation is the fastest ancillary service, responding within seconds to maintain system frequency. Responsive reserve is online and responds within ten minutes. Non-spinning reserve responds within thirty minutes. The quantity and types of ancillary services required for reliability depend on features of the electricity system, such as the mix of resources and the largest contingency.

To maintain the system frequency of 60 Hertz, load frequency control is run every 4 seconds. Load frequency control sends signals for regulation up, regulation down, and responsive reserve to restore system frequency.

When a large generator trips, regulation up attempts to address the immediate frequency drop, while responsive reserves ramp up and non-spinning reserves are brought online. Because of inertia, the failure of even a large unit causes a gradual decline that must be made up by the reserves. The system is much like a bicycle. When the bicyclist stops pedaling, the bicycle gradually slows down. To maintain the bike speed, all that is needed is another to start pedaling with the same force. In fact, a large electricity system is like a large bicycle with hundreds pedaling. Inertia from the synchronized thermal resources keeps the system running even in response to large shocks.

Reserves are tightly woven into the spot market. The day-ahead market jointly optimizes the purchase of energy and reserves. Most US markets also co-optimize energy and reserves in real time, although ERCOT has yet to implement this enhancement. Real-time co-optimization of energy and reserves further enhances the selection and pricing of reserves throughout the operating day.

Scarcity pricing

Sending the right real-time price signal is critical to motivate efficient behavior in real time, as well as further forward decisions including long-term investment. In normal times, this price signal follows from the marginal cost of supply or the marginal value of demand. However, in instances of scarcity where the system operator has limited reserves to maintain power balance, the value of the reserves—and the price of energy—should reflect the value real-time reserves create in avoiding load-shedding events. The marginal value of reserves is equal to the value of lost load (VOLL) in extreme shortage situations and then falls as the scarcity is less and therefore the probability of lost load is less. ERCOT employs an operating reserve demand curve to reflect the marginal value of reserves in scarcity (Hogan 2013). As the availability of reserves falls the marginal value of reserves increases at an increasing rate until it reaches VOLL in severe shortages, as shown in Figure 1. VOLL is set administratively at $9000/MWh—367 times higher than the average energy price of $24.62/MWh in 2016. During scarcity events, the operating reserve demand curve sets the price of online reserves. The energy price also includes this price as an adder.

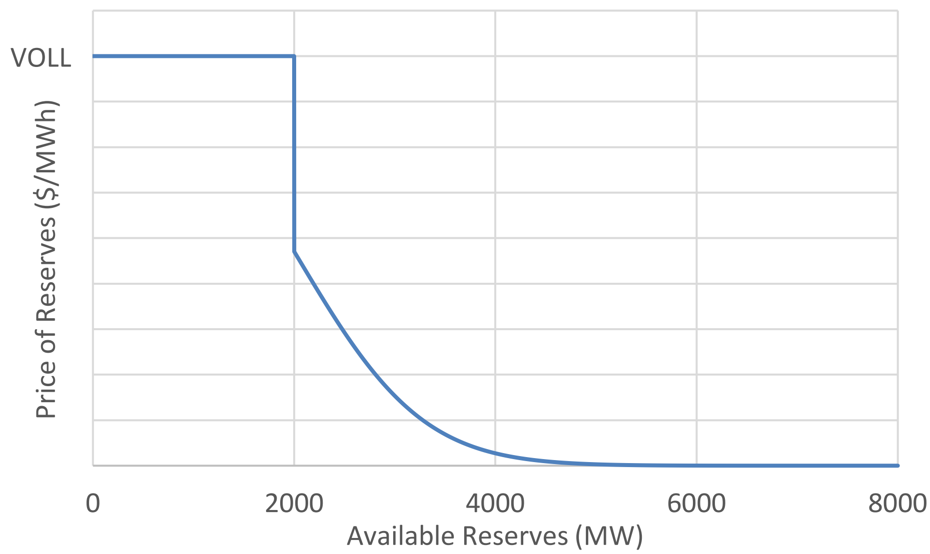

Figure 1: ERCOT's operating reserve demand curve (source: ERCOT)

Scarcity pricing is the administrative approach used to implicitly express load's preference for reliability. The higher the scarcity pricing, the stronger is the investment signal. More generation is economically built, which improves reliability. The higher scarcity pricing also motivates demand response to avoid the high prices. Indeed, it is the limited amount of demand response that creates the need for scarcity pricing.

Without scarcity pricing, reserve and energy prices are apt to be too low during the critical periods of scarcity, because of rules that mitigate market power. Relying on the exercise of market power on the supply side to push prices higher during scarcity is a poor solution. With scarcity pricing a generator can offer at marginal cost, be efficiently dispatched in real time, and still receive a price that reflects its value in scarcity. Scarcity pricing also allows the system operator to co-optimize energy and reserves in real time, as the marginal value of reserves in avoiding load-shedding events is properly considered.

Forward contracts

As we have seen, the core components of the market design are the day-ahead and real-time markets. These spot markets provide the basis for efficient scheduling of resources and then real-time dispatch based on the latest information. The spot markets also provide a framework for forward contracting. Both markets let participants report forward trades, which are then settled by the system operator.

Forward contracts enable market participants to better manage risk. The ability to manage risk is essential to market stability. Markets fail when large participants default. The California energy crisis of 2000-2001 is a vivid example. In California, poor market rules prevented the large utilities from engaging in forward contracting. These utilities had obligations to serve load at fixed rates yet faced potentially volatile spot prices. When an extended period of shortage caused spot prices to spike in California, the unhedged utilities quickly became insolvent and the market collapsed. With forward contracting, the utilities could purchase energy in advance consistent with their obligations.

An important advantage of scarcity pricing is that it motivates load to contract for the energy it needs in advance of real time. Forward contracting provides a hedge against volatile real-time prices. Forward contracting thus reduces price risk. Load is hedged by the forward contract with supply, and supply covers its forward position with its generation and fuel contract. In this way, overall risk in the market is reduced. All parties have some form of hedge, either financial or physical.

Forward contracting has a second advantage. It improves bidding incentives. There are two parts to the argument. First, to minimize risk, participants have an incentive to take a forward position in the market that is consistent with eventual needs. The reason is that prices become more volatile as we get closer to real time and the options for balancing the market become more limited, so supply and demand curves are steeper. For example, the extremely high scarcity prices of up to $9000/MWh, only arise in real time as reserves become scarce. Forward trade consistent with eventual needs avoids these more volatile prices. Second, market participants entering the real-time market in more balanced positions have less incentive to exercise market power. In a clearing-price auction, sellers have an incentive to offer above marginal cost and buyers have an incentive to bid below marginal value (Ausubel et al. 2014). However, a participant that enters the real-time market in a balanced position has no incentive to overstate or understate. Forward contracting induces participants to submit bids and offers that are more consistent with true preferences, improving the efficiency of the market outcome (Allaz and Vila 1993, Cramton 2004).

Practice confirms these benefits of forward contracting. The natural buyers—those with load obligations—and the natural sellers—the generators—have a strong tendency to engage in trade to manage risk. Typically, most energy is contracted well in advance of the day-ahead market in monthly, annual, or multi-year forward contracts, and even less is traded at the more volatile real-time prices. This does not diminish the importance of the day-ahead and real-time markets as they provide the foundation on which forward trade is based. Forward trade anticipates the expected prices in the day-ahead market in the same way the day-ahead market anticipates the expected prices in the real-time market. Arbitrage across markets drives forward prices toward the expected prices observed closer to real time. Moreover, the day-ahead and real-time markets provide an efficient means, especially for smaller players, to adjust forward positions. This is important, because it mitigates an advantage a dominant incumbent would otherwise have from its control of a large fleet of resources.

The benefits of forward contracting are reinforced with more transparent forward markets. This is one advantage of having the system operator settle bilateral trades. Although the terms of trade are not public, they are observable by the system operator and the market monitor. In some markets, transparent auctions are conducted to establish retail pricing for customers who do not select a retail provider. Transparent forward prices can motivate entry into the retail market and strengthen both retail and wholesale competition.

On the other hand, absent an efficient and liquid spot market, forward contracting can be a means for dominant incumbents to discourage entry and limit competition, as seems to be the case in Great Britain (Newbery 2005 and Helm 2015). However, the source of the problem in Britain appears to be a poor spot market, rather than forward contracting.

Congestion Revenue Rights

One forward market that is organized by the system operator is the market for congestion revenue rights. The system operator conducts semi-annual and monthly auctions for point-to-point congestion revenue rights (CRR). These auctions enable participants to hedge congestion risk. A generator at node X that sells energy to load at node Y can buy a X-to-Y CRR and eliminate congestion risk. The holder of the CRR is paid the day-ahead congestion revenue from source-to-sink. Although there are a vast number of CRR products, in practice the problem is manageable because only a small subset of lines experience significant congestion. It is this more limited set that are of economic significance and are traded to manage congestion risk.

Capacity market

Some electricity markets, such as ISO New England and PJM, include a long-run investment market for capacity. The capacity market coordinates new investment and retirements to assure that adequate resources are available to reliably meet load. Other markets, such as ERCOT, are energy-only markets. An energy-only market relies solely on the price signals from the day-ahead and real-time markets to induce sufficient investment in resources to reliably meet load. The challenge with an energy-only market is that it typically takes several years to build new generation. If resources become inadequate, there may be few good options to address the shortfall in the time available. A capacity market can serve as an insurance policy and coordination device to better optimize investments, both in time and location.

Capacity markets initially were poorly designed (Cramton 2003). The early markets provided a means for generators to secure some additional revenues, but offered little value to load. There were two key problems. First, the early markets were organized as spot markets, but capacity costs are almost entirely sunk in the short-term, so that the supply curve is 0 for existing capacity and then essentially vertical. On the other side of the market, demand is a fixed requirement imposed on load. Thus, price formation is quite poor, since the market is attempting to cross a vertical demand curve with a vertical supply curve. The clearing price is either 0 or the price cap and the incentive to exercise market power is severe. Second, early markets defined the product poorly based primarily on nameplate capacity, rather than on something related to reliability. Because of these flaws, early capacity markets did little more than transfer a rather arbitrary amount of money from load to generation.

Modern capacity markets, such as those run by ISO New England and PJM, correct the two basic flaws (Cramton et al. 2013). First, the capacity market is conducted several years in advance of delivery. This means that new entry can compete with existing supply and the cost of new entry can set the price in growing markets. This greatly improves price formation. Second, the product is properly defined as the ability to generate energy during scarcity. Energy during scarcity is the product that contributes to reliability. With this approach, load is buying an option for energy during scarcity periods. The product is both physical and financial. Resources can only offer capacity consistent with their physical capability to provide energy during scarcity. This guarantees there are adequate resources to meet system peaks. The financial component is a call option for energy during periods of reserve shortage, when scarcity pricing kicks in. A supplier's obligation to provide energy during shortages follows load; that is, a supplier that has a 10% share of the capacity purchased by load has an obligation to provide 10% of the energy during shortages. Deviations are priced at the scarcity price, which is over $5000/MWh in New England.

The capacity product defined in this way fully hedges load from scarcity prices. Moreover, it maintains the strong performance incentives for generators, who see and feel the high scarcity price on the margin that impacts decisions, a feature called pay-for-performance. A generator that provides more than its share is paid the scarcity price for each MW over and is charged the scarcity price for each MW under. In addition, the product provides all the benefits of forward contracting. Risk is reduced, since load is fully hedged and supply has its capacity and fuel contracts as a physical hedge. And the obligation puts the generator in a roughly balanced position entering the real-time market, greatly reducing incentives to exercise market power under scarcity conditions. Suppliers receive a constant capacity payment across all hours instead of the highly volatile scarcity rents they would receive in an energy-only market without forward contracting.

Modern capacity markets, based on pay-for-performance, are closely related to the energy-only market in Texas with high scarcity pricing and extensive forward contracting. In Texas, the high scarcity pricing motivates the forward contracting that limits risk and induces investment. The scarcity price is the key instrument for resource adequacy. One reason this may work well in Texas is substantial industrial load that makes the market for forward contracts more liquid. Perhaps even more important is the attractive investment environment in Texas with abundant fossil, wind, and solar resources, lots of land, lots of transmission, and a friendly regulatory environment.

The key feature that modern capacity markets add to the energy-only model is the capacity requirement that guarantees adequate physical resources. An annual auction is conducted that procures resources several years in advance to satisfy the capacity requirement. Importantly, the requirements respect transmission limits. A load pocket with too few resources can have a higher capacity price to induce entry in that location. The outcome of the capacity auction not only assures adequate resources overall, but in each load zone. By contrast, the energy-only market in Texas relies solely on the energy LMPs for locational incentives and these may be too weak. Especially when efficient investments require large, lumpy resources, building the resource in the high LMP location may destroy the congestion rents that would otherwise attract it. Long-term power purchase agreements may mitigate this problem, but not without buyer concentration in the load pocket, so that the buyer can internalize the benefit of the forward purchase.

The main advantage of capacity markets is the coordination of investment in resources and a stronger guarantee of adequate resources. The markets are however difficult to design and operate. Harbord and Pagnozzi (2014), Newbery and Grubb (2015), and Newbery (2016) discuss some of the issues. One thing is clear. The case for capacity markets is stronger in regions with greater investment challenges and more locational constraints. However, both capacity and energy-only market models depend critically on well-functioning spot markets and a stable regulatory framework to promote efficient investment.

Governance

Some markets operate well with few changes over many decades. Electricity markets are not among these markets. Electricity markets are much too complex for the market design to stand still. Successful electricity market designers are constantly looking ahead to address challenges and exploit opportunities as circumstances change. Good governance is essential in making sure the market design evolves in desirable ways.

Fortunately, there are some successful models of governance. One of the blueprints for governance that appears to work well is the one applied in the Texas electricity market. The market has succeeded for twenty years, encouraging both wholesale and retail competition in a reliable electricity market.

There are several layers of governance: the regulator, the board, the participant committees, the market monitor, and the system operator.

Electricity is a regulated industry, so a government commission has primary jurisdiction over the system operator's activities. This is the energy regulator of the jurisdiction, for example the Federal Energy Regulatory Commission in most of the US or the Public Utility Commission in Texas. The regulator establishes the mission of the system operator and the broad principles of the market design. The regulator provides oversight, but does not manage the details. Matters that cannot be resolved by the system operator or its board can be appealed to the regulator.

In Texas, the system operator is governed by a board of directors made up of independent members as well as representatives from all stakeholder groups. Including stakeholder representatives on the board has the advantage of direct representation of stakeholder interests and knowledge in a transparent way. The board appoints the system operator's officers to manage day-to-day operations consistent with the system operator's mission. The board reviews and adopts any changes to the market rules.

A technical advisory committee studies market rule changes and makes recommendations to the board of directors. The committee has broad stakeholder representation and is assisted by workgroups focused on technical matters.

An independent market monitor reports regularly to the board. The market monitor is independent of the system operator management. The job of the market monitor is to observe the market, identify problems, and suggest improvements. Importantly, the market monitor is not a judge or decision-maker; the market monitor is an independent expert voice. This allows the market monitor to react to problems more quickly than the regulator. Independent market monitors were introduced in all US electricity markets following the California energy crisis of 2000-2001.

The Independent System Operator (ISO) runs the market. The ISO acts as a trusted party, ensuring transparency, fairness and efficiency. The ISO is independent of both the demand and supply side of the market. The ISO is guided by a mission, which in Texas is: "To serve the public by ensuring a reliable grid, efficient electricity markets, open access and retail choice."

The ISO operates with a high level of transparency. Market improvements are discussed and developed with participation of stakeholders. The ISO plays a lead role in identifying problems and developing solutions. This is essential. Complex markets are best designed by experts, not committees. The technical advisory committee plays an important role in discussing proposed changes and debating alternatives. Then the board with the oversight and guidance from the regulatory commission decides on any changes to the market rules. In ERCOT, 90 rule changes were approved in 2016. Many of the changes are the result of technological progress, such as accommodating new resource types.

This governance structure appears to be working well in promoting the mission of the market. There has been a steady improvement in the market rules. Yet despite the changes, there is a stable underlying framework. This framework provides the basis for the robust long-run investment seen in the market. Most importantly, there is no evidence of regulatory capture, which is a common problem in regulated industries. Captured markets are typified by low levels of competition and innovation and high profits. In Texas, both the wholesale and retail markets are highly competitive, profits are low, yet innovative new technologies are being introduced.

Regional differences

So far I have focused on the predominant market design in North America. That design features central optimization of resources both day-ahead (for unit commitment) and in real time (for security-constrained economic dispatch) and locational marginal prices to reflect binding transmission constraints.

In Europe, electricity markets remain more fragmented than in the US. This fragmentation stems from national markets with limited integration across countries. In the US, the system operators cover larger regions, and within these large regions full optimization of resources occurs both day-ahead and in real time. Locational price signals are much weaker in Europe—typically there is a single price of energy for each country. Transmission congestion both within country and across countries is not efficiently priced. Although there are plans for much tighter integration, severe challenges remain given that market designs are still developed at the national level. The European Commission is providing some guidance to member states, but the guidance tends to be about broad principles, rather than the detailed market rules required for tight integration. Improvements are certainly being made, but I suspect fully optimized commitment and dispatch of resources together with efficient congestion pricing would reap large gains, especially with the rapidly increasing penetration of renewable resources, which put considerable pressure on the transmission grid.

European markets include a voluntary balancing market in real time instead of the security-constrained economic dispatch of the US. The balancing market tends to be thinner than the real-time markets in the US, which can lead to less reliable real-time prices. This creates greater demand for intraday trading to resolve imbalances in advance of real time. Traders certainly like this system and indeed lobbied for a model of continuous trading familiar from financial exchanges. The model, however, is problematic as it encourages an arms race for speed (Budish et al. 2015). Continuous trading is unworkable in electricity markets, because combinatorial optimization is required to establish the feasibility of trade and this requires significant computation time (on the order of minutes). Thus, continuous trading results in unacceptable backlogs in order processing. A more sensible approach is intraday auctions, say every fifteen minutes, that would allow sufficient time for computation. There appears to be a move in this direction (Neuhoff et al. 2016).

The problem of a thin balancing market is aggravated in Great Britain with an odd pricing rule. Rather than settle at the clearing price, in the balancing market of Great Britain both demand and supply are pay-as-bid. With this approach, the area between the aggregate demand and supply curves (the *as-bid* gains from trade) is kept by the system operator. This acts as a tax to trade in the balancing market, discouraging its use. This was intended to encourage forward contracting and perhaps mitigate market power, but it has a negative side-effect. The tax is greatest for smaller players who are less able to guess the clearing price in the balancing market. Thus, the pricing rule favors greater concentration on both the supply and demand sides. The greater concentration then frustrates liquidity in the forward markets. Strong incumbent positions are reinforced.

Differences in market designs often arise from important differences in regions as well as the historical evolution from monopoly utilities. The Nord Pool market in Northern Europe is a good example. This market began in Norway, a hydro-dominated system with vast amounts of reservoir storage and relatively affluent customers. Such a setting is well-suited to the power exchange model based on day-ahead and intraday trading. The complexities of fully-optimized unit commitments and security-constrained economic dispatch are not needed. The economic problem in Nord Pool is made much simpler because of the flexible hydro resources that effectively store and release energy based on opportunity cost. Thermal systems without significant storage, as in the US, must grapple with the non-convexities of startup and minimum-energy costs as well as other limits in flexibility. The success of the Nord Pool market led to its expansion in Northern Europe. Nord Pool has played a large role in steering market designs in Europe toward a power exchange market based on day-ahead and intraday trading. It remains to be seen whether this simpler market design extends well across Europe.

In South America, Brazil provides a sharp contrast with Norway and Nord Pool. Like Norway, Brazil is hydro-dominated, but there are important differences: (1) Brazil is relatively poor, (2) Brazil is growing rapidly, and (3) Brazil's periods of extremely dry weather are more extreme and rarer (roughly once in ten years) than in Norway. Combined, these features mean that Brazil needs lots of new investment including thermal resources that are only rarely used to provide energy during the extreme dry periods. Brazilians also cannot accept sustained high prices during these dry periods. Because of these challenges, the Brazilian market is focused on planning and competitive procurement of new resources under long-term power-purchase agreements. Once procured, the resources operate in a centrally optimized day-ahead and real-time market, much like in the US, but the offers are cost-based, as determined by administrative formulas, rather than freely bid. The opportunity cost offers of the hydro-resources are also determined administratively from dynamic optimization. Brazil effectively moved the competitive bids to the investment stage, and then relies on central optimization to assure that it gets the most from the procured resources. This model does address the investment challenges, although it has its own flaws (Dutra and Menezes 2017).

Colombia is similar in setting to Brazil. Colombia, however, operates a bid-based market more like the US. One difference with the US is Colombia is a single-price market without LMPs. For this to work well, Colombia must build ample transmission as the market grows to avoid the binding transmission constraints that otherwise can unravel a single-price market. To address investment, Colombia has a firm energy market like the modern capacity markets in the US (Cramton and Stoft 2007). As in the US, the product is an obligation to generate energy during scarcity, but in Colombia the scarcity event is an extreme dry season, rather than a capacity shortage.

Key issues looking forward

Electricity markets are undergoing a radical transformation. The transformation is primarily driven by a move away from fossil fuels to address climate change and reduce local pollution. Markets are experiencing a rapid increase in wind and solar. In the US, 60% of new capacity in 2016 was from solar and wind (US EIA 2017). This trend is expected to continue until fossil fuels, especially coal, are largely eliminated from the energy mix. A second driver of change is technological progress. Some of this progress is seen in the falling cost of wind and solar. But there are also advances in storage, distributed generation, as well as technologies that allow greater demand response, such as smart homes. Some of these changes present new market challenges; others offer solutions.

First consider the expansion of renewables. To date, this expansion has largely come from wind, because it has been less costly, but solar cost is dropping rapidly, and soon will be the least-cost new resource in many regions. Solar and wind share three characteristics that pose new challenges for the market: (1) intermittent supply, (2) zero marginal cost, and (3) no (or limited) inertia. Today's electricity systems can and do handle a moderate share of solar and wind. This is because there are sufficient other resources to step in when wind and solar output is low and there is sufficient inertia from the thermal resources. But imagine a system with nearly all energy coming from wind and solar. The system could not work without other elements. Without inertia, the system would require some means of instantaneous response to balance supply and demand. As inertia falls and supply uncertainty increases more reserves and faster responding reserves are required.

Battery storage is one approach to address the challenges of renewable energy. Batteries are well-suited to store energy during periods of peak supply and then release it when the energy is of greatest value. Batteries also respond nearly instantaneously, addressing the problem with inertia. Battery cost is falling rapidly as the market expands in part from the large-scale introduction of electric vehicles.

Demand response, especially from smart home technologies, also will help address the challenges of renewable energy. As more and more appliances become computer controllable, greater opportunities for time-shifting of demand or demand curtailment are possible. Electric vehicles present an opportunity for combining battery storage and demand response. Charging batteries at night is a powerful solution: the wind typically is stronger, other electricity demand is lower, and time-shifting is costless and easily managed. Moreover, large penetration of solar will shift the net peak to the evening (to around 8 pm). Discharging EV batteries during the evening net-peak (net of solar) will meet demand and provide the rapid ramping that is required as the sun sets.

But will the market provide the investment incentives for battery storage and improved demand response as the share of renewable resources grows? I believe the answer is yes. With insufficient flexible resources, energy prices will become highly volatile, especially in real time. Flexible batteries, demand response, and gas peakers are well positioned to profit from this volatility, buying when prices are low and selling when prices are high. Scarcity pricing—high pricing during times of scarcity—is essential in strengthening this investment signal.

Today's best-practice market designs likely can support high levels of renewables. Markets in Texas and Europe have run reliably with more than 50% of energy coming from wind. What is required as renewable penetration increases is the routine adjustment of market parameters such as the ancillary services requirements, as well as the steady improvement in the market design to make the system more responsive and flexible. Planning plays a critical role as well in identifying problems early.

One issue that frustrates market success, especially with respect to long-run investment, is an incoherent and unstable climate policy. Subsidies should be limited to research and the early deployment of emerging technologies where learning-by-doing cost reductions are best thought of as public goods. A climate policy built on a myriad of changing subsidies and emission restrictions makes planning much more difficult both for the system operator and investors. The resulting uncertainty harms the investment environment. By contrast a coherent and stable policy based on a carbon price would greatly reduce investment uncertainty (Cramton et al. 2017). For any investor making a 30-year investment in electricity, a critical input in the decision is the carbon price path going forward. Energy companies today assume some carbon price to identify the best investments. Less uncertainty about the carbon price reduces risk and improves decision making. Climate policy, although beyond the scope of electricity market design, is something of first-order importance that the market must contend with. An unstable climate policy both weakens and distorts investment incentives.

Which market design is best?

Before concluding, let me return to Wilson's (2002) discussion of the two alternative market architectures: the *integrated model* in which a system operator centrally optimizes resources to maximize welfare subject to constraints and the *exchange model* in which energy and related products are traded day-ahead and throughout the day at prices that clear the market without the benefit of central optimization. At the time Wilson was writing his Presidential Address in the late 1990s, the debate between market models was raging in the US. The East favored the integrated model and the West favored the exchange model. Both market models at that time had severe flaws. Yet a case could be made that the simpler exchange model might perform better as traders could more nimbly react to market forces, rather than being constrained to operate within a highly flawed central optimization. This indeed is the core critique of centralization. The system operator is maximizing the wrong thing: the objective is wrong, the constraints are wrong, and the bids are wrong. Despite this possibility, in the US, the integrated model carried the day. The transition to the integrated model in the West was hastened by the dramatic failure of the exchange model in California in 2000-2001.

Much has changed in the last twenty years. Importantly, our ability to model and optimize large networks has advanced by many orders of magnitudes. The system operator now optimizes much more realistic models of the network and resources. State-of-the-art integrated systems now co-optimize energy and reserves both day ahead and in real time. Real time dispatch need not be a static optimization, but can look forward to better handle dynamic constraints. Uncertainties can be explicitly modeled. Beyond these technical advances, the integrated systems have seen major improvement in the market rules to address potential failures such as market power and to include innovations such as demand response. Here is where good governance is essential in making sure improvements are made in a timely way. To me, these advances over the last twenty years decidedly shift the scales in favor of the integrated model.

The reality of electricity is that system and resource constraints must be satisfied to keep the lights on and avoid the destruction of equipment. This is true with either architecture. Today's integrated model simply does a better job of addressing and pricing the constraints via direct optimization. Efficient transmission pricing is a lead example. The exchange model operating in much of Europe is improving, but still falls short in pricing transmission congestion both within and across countries.

In some sense, the dichotomy between the two architectures is false. Both are hybrids in practice. Today's integrated model includes basic elements of the exchange model. Participants can submit virtual bids and offers in the day-ahead market that are not associated with any physical resource. As discussed earlier, these purely finance offers have improved market outcomes. One subtle but important benefit of virtual bidding is the inclusion of a large quantity of convex bids and offers. This creates a more convex optimization. Pricing anomalies arising from non-convexities become less important. Integrated markets also allow self-scheduling of resources. A generator has the option of simply telling the system operator how a resource will run, rather than expressing the unit's economics and letting the schedule follow from the system operator's optimization. Self-scheduling still involves optimization, but the optimization is done by the generator not the system operator. The generator becomes a price taker and is only asked to deviate from the self-schedule in extreme circumstances. Finally, forward trading is a key element of both architectures. The integrated model does not reduce the need for forward trading, rather it supports forward trading through robust day-ahead and real-time prices.

With respect to bidding incentives, neither architecture is perfect. Still I have become convinced that market designs that provide good incentives for truthful bidding in favorable conditions can be made to work reasonably well in difficult settings. The integrated model works perfectly in competitive environments without non-convexities; so too does the exchange model when markets are complete. The integrated model, however, better handles non-convexities and is simpler for participants. A good example of this is the use of three-part offers, as mentioned before. This lets the generator offer true costs and puts the burden of optimizing the unit's schedule on the system operator armed with much more complete knowledge of the system and the other resources. Also, market power is more easily mitigated with three-part offers, since deviations from cost-based offers are more transparent. With the one-part offers of an exchange, distortions from marginal cost can be justified from startup and minimum-energy costs.

An important advantage of the integrated model is competition. A dominant incumbent with a large and diverse fleet of resources can do a good job of optimizing the use of the fleet. The dominant incumbent gains less from the system operator's optimization of the entire system. Small players, by contrast, stand to gain much more from the system optimization, both the scheduling of resources day ahead and then the efficient settlement of deviations in real time. The integrated model provides opportunities and protections for the smaller generators that may be missing in the exchange model. The absence of a liquid real-time market in Europe has raised competition concerns both in Great Britain and on the continent.

The integrated model also supports competition through transparency. The system operators in North America have a remarkable degree of transparency. The mission, the governance, the market rules, the market data, all are highly transparent. The exchange-based markets also have a high level of transparency, but in my experience, they tend to be less transparent, perhaps because traders and large incumbents desire to conceal their manipulations.

One can argue that the integrated model may lack innovation. Centralization can lead to bureaucracy that stifles innovation. This again is where good governance must come to the rescue. Improved competition, especially enabling small participants with good ideas, is one important source of innovation, but it is also essential that the voice of innovators be heard. Innovative improvements to market rules must be encouraged. The most successful markets in the long-run will be those that foster innovation consistent with the market objectives.

The experience from financial markets would suggest that the exchange model does not necessarily get high marks for innovation, even in a competitive environment. Financial markets provide a good example of the exchange model in a highly competitive environment. Financial markets certainly have improved over the last twenty years with the adoption of information technology that has greatly reduced transaction costs. However, over the last decade, the innovation largely has been on speed—trades that once took seconds now take microseconds. These speed gains have benefited fast traders and the exchanges able to sell speed-enabling tools, but not fundamental investors. The innovation has taken the wrong form because of a basic flaw in market design that rewards speed (Budish et al. 2015). By contrast, in integrated electricity markets the innovation has been on improved methods of finding gains from trade through improved optimization and better market rules. The contrast between integrated electricity markets and financial markets is extreme. To accommodate speed, today's financial markets avoid all computation at the exchange. This limits how preferences are expressed and gains from trade are found.

The integrated model also is readily customized to address the challenges of the setting, and to evolve as those challenges change. It is not one size fits all. The market may rely solely on scarcity pricing to address reliability or it may combine scarcity pricing with a capacity market to further assure adequate resources. The collection of ancillary services and quantities needed also are tuned to the setting.

I do believe that there are settings where the exchange model can work well. Nord Pool’s hydro-dominated system is the leading case. Because of ample flexible capacity and large amounts of energy storage, the simpler exchange model works well. Non-convexities are relatively unimportant.

However, in most other settings, the integrated model has compelling advantages. System-wide direct optimization of resources subject to constraints simply does a better job of identifying and realizing gains from trade, as well as establishing prices that reflect system conditions. These prices guide behavior from real-time performance to long-term investment.

Conclusion

Electricity markets offer a fascinating window into the practice of market design. Electricity market design must draw on expertise from economics, operations research, engineering, and computer science. Despite the complex design problem, the progress toward successful market designs has been rapid. One reason for this success is a feature of electricity. Blackouts and excessive prices are politically unacceptable. This imposes some discipline on the evolution of the markets. The fact that electricity networks must satisfy rigid constraints to avoid destruction reinforces this discipline. The markets simply must work reasonably well and there are strong incentives for improvement.

The successful market design in the US has at its core the day-ahead market for unit commitment and scheduling, followed by security-constrained economic dispatch in real time. The market allows direct expression of a resource's underlying economics and constraints, then joint optimization of all resources to maximize social welfare. The market relies on linear prices, locational marginal prices, that establish the marginal value of energy at each time and location given the network constraints. This spot market provides the basis for forward contracting, which allows market participants to manage risk. Market power is a potential source of market failure, especially as we get closer to real time, and so the rules include automatic methods of mitigation—such as substituting cost-based bids when market power is excessive. Still, extremely high prices during periods of scarcity play an essential role in motivating real-time performance and the investments needed for reliability.

Looking forward, electricity market design will continue to be an exciting area. The transformation of resources from fossil fuels to renewables raises non-trivial issues. As supply uncertainty increases and inertia falls, the electricity system must become increasingly responsive. New storage technologies and demand response from smart homes may provide a solution, but this will require some market rule enhancements.

Despite these technological changes, the best market designs for the future will continue to rely on a highly efficient spot market with strong support for forward contracting, as well as a competitive retail market to foster innovative demand response. This core framework is best apt to support efficient long-run investment, the key goal of the restructured markets. A stable and coherent climate policy based on a carbon price would further support this goal by greatly reducing investment uncertainty.

References

Allaz, Blaise and Jean-Luc Vila (1993), "Cournot Competition, Forward Markets and Efficiency," *Journal of Economic Theory*, 59: 1-16.

Ausubel, Lawrence M., Peter Cramton, Marek Pycia, Marzena Rostek, and Marek Weretka (2014), "Demand Reduction and Inefficiency in Multi-Unit Auctions," *Review of Economic Studies*, 81:4, 1366-1400.

Borenstein, Severin (2002), "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster," *Journal of Economic Perspectives*, 16:1, 191-211.

Budish, Eric, Peter Cramton, and John Shim (2015), "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics*, 130:4, 1547–1621, November.

Chao, Hung-Po, Shmuel Oren, and Robert Wilson (2008), "Reevaluation of Vertical Integration and Unbundling in Restructured Electricity Markets," in Fereidoon P. Sioshansi, ed., *Competitive Electricity Markets*, Chapter 1.

Competition and Markets Authority (2016), "Energy Market Investigation: Summary of Final Report," June.

Cramton, Peter (2003), "Electricity Market Design: The Good, the Bad, and the Ugly," *Proceedings of the Hawaii International Conference on System Sciences*, January.

Cramton, Peter (2004), "Competitive Bidding Behavior in Uniform-Price Auction Markets," *Proceedings of the Hawaii International Conference on System Sciences*, January.

Cramton, Peter, David JC MacKay, Axel Ockenfels and Steven Stoft (2017), *Global Carbon Pricing—The Path to Climate Cooperation*, MIT Press.

Cramton, Peter, Axel Ockenfels, and Steven Stoft (2013), "Capacity Market Fundamentals," *Economics of Energy & Environmental Policy*, 2:2, September.

Cramton, Peter and Steven Stoft (2007), "Colombia Firm Energy Market," *Proceedings of the Hawaii International Conference on System Sciences*, January

Dutra, Joisa and Flavio M. Menezes (2017), "Electricity Market Design in Brazil: An Assessment of the 2004 Reform," Working Paper, University of Queensland.

ERCOT (2017), *Inside the Promise: 2016 State of the Grid Report*, February.

Federal Energy Regulatory Commission (2002), "Working Paper on Standardized Transmission Service and Wholesale Electric Market Design," March.

Harbord, David and Marco Pagnozzi (2014), "Britain's Electricity Capacity Auctions: Lessons from Colombia and New England," *Electricity Journal*, 27:5, 54-62.

Helm, Dieter (2015) "Competition in the British Electricity Sector: A Set of Practical Reforms," *Energy Futures Network*.

Hogan, William W. (2013), "Electricity Scarcity Pricing Through Operating Reserves," *Economics of Energy and Environmental Policy*, 2:2, 65-86.

Hogan, William W. (2014), "Electricity Market Design and Efficient Pricing: Applications for New England and Beyond," *Electricity Journal*, 27:7, August.

Jha, Akshaya and Frank A. Wolak (2014), "Testing for Market Efficiency with Transaction Costs: An Application to Convergence Bidding in Wholesale Electricity Markets," Working Paper, Stanford University.

Neuhoff, Karsten, Nolan Ritter, Aymen Salah-Abou-El-Enien, and Philippe Vassilopoulos (2016), "Intraday Markets for Power: Discretizing the Continuous Trading?" Discussion Paper, DIW Berlin.

Newbery, David (2005), "Electricity Liberalisation in Britain: The Quest for a Satisfactory Wholesale Market Design," *Energy Journal*, 43-70.

Newbery, David (2016), "Missing Money and Missing Markets: Reliability, Capacity Auctions and Interconnectors," *Energy Policy*, 94, 401-410.

Newbery, David and Michael Grubb (2015), "Security of Supply, the Role of Interconnectors and Option Values: Insights from the GB Capacity Auction," *Economics of Energy and Environmental Policy*, 4:2, 65-81.

O'Connor, Phillip R. and Erin M. O'Connell-Diaz (2015), "Evolution of the Revolution: The Sustained Success of Retail Electricity Competition," *COMPETE Discussion Paper*.

Stoft, Steven (2002), *Power System Economics*, IEEE Press.

US Energy Information Administration (2017), "US Electric Generating Capacity Increase in 2016 Was Largest Net Change Since 2011," *Today in Energy*, February 27.

Wilson, Chris M. and Catherine Waddams Price (2010), "Do Customers Switch to the Best Supplier?" *Oxford Economic Papers*, 4, 647-668.

Wilson, Robert (2002), "Architecture of Power Markets," *Econometrica*, 70:4, 1299-1340.