RAP®

REGULATORY ASSISTANCE PROJECT

NOVEMBER 2020

Demand Charges: What Are They Good For?

An Examination of Cost Causation

Mark LeBel and Frederick Weston, with contributions from Ronny Sandoval

Contents

Acknowledgments

Editorial assistance was provided by Ruth Hare and Donna Brutkoski.

The authors would like to acknowledge and express appreciation to the following people who provided helpful insights into early drafts of this paper:

- Janet Gail Besser, Smart Electric Power Alliance
- Paul Chernick, Resource Insight
- Dan Cross-Call and Becky Li, Rocky Mountain Institute
- Karl R. Rábago, Rábago Energy LLC
- Jim Lazar, Carl Linvill and Ann McCabe, RAP

We would also like to thank GridLab for providing funding for Ronny Sandoval's participation in this project. When the report was drafted, Mr. Sandoval was working as an independent consultant. At the time of publication, he is employed at Vote Solar.

The analysis in this paper rests on the shoulders of earlier RAP publications, in particular:

Lazar, J., & Gonzalez, W. (2015). *Smart rate design for a smart future*. Regulatory Assistance Project. https://www.raponline.org/knowledge-center/smart-rate-design-for-a-smart-future/

Linvill, C., Lazar, J., Dupuy, M., Shipley, J., & Brutkoski, D. (2017). *Smart non-residential rate design*. Regulatory Assistance Project. https://www.raponline.org/knowledge-center/smart-non-residential-rate-design/

Weston, F. (2000). *Charging for distribution utility services: Issues in rate design*. Regulatory Assistance Project. https://www.raponline.org/knowledge-center/charging-for-distribution-utility-services-issues-in-rate-design/

That said, responsibility for the information and views set out herein lies entirely with the authors.

Executive Summary

Demand charges, rates that are applied to an individual customer's maximum short-term usage (typically 15, 30 or 60 minutes) in a billing period, have existed since nearly the beginning of the electric industry. While utilities often favor demand charges, economists have continually questioned whether they are an efficient form of pricing. With the widespread adoption of advanced metering, this is an opportune time to reconsider demand charges, even for industrial customers, and replace them with more efficient time-varying energy (kilowatt-hour) rates.

Traditional monthly demand charges have always provided a perverse incentive that does not reflect cost causation for shared system costs. Individual customer noncoincident peaks (NCPs) do not reflect the coincident peaks that drive *shared* generation and delivery capacity costs. The price signal that demand charges send — to lower individual customer NCP and to level a customer's load over time — is substantially different than a price signal to reduce usage at the time of coincident peaks. As a result, demand charges penalize customers for usage at times that do not impose particularly high costs and encourage them to waste effort and money shifting loads off their own maximum hour (and sometimes onto high-load system hours).

The historic exception to this rule is a customer that has a nearly 100% coincidence factor with the relevant peaks. The prototypical example of this in the mid-20th century was an industrial customer with very high load factors. Demand charges could be reasonable in the past only as applied to this specific category of customers. But, in today's electric system, even this justification for demand charges falls away. High penetrations of nondispatchable but variable renewable generation means that a 100% load factor is unlikely to be, from a system perspective, the most desirable load shape. Rather, flexible load — load that can respond to swift changes in the availability of supply, perhaps in the middle of the day for solar and late at night with wind — becomes cheaper to serve than unvarying loads in systems marked by high penetrations of variable supply.

Historically, demand charges have frequently been sized to recover most or all shared system capacity costs. Again, this may have been reasonable enough in the mid-20th century for certain customers, but it does not reflect the economics and engineering of a modern electric system. The choices that system planners make are trade-offs between different types of costs. Much "capacity" investment today aims to reduce energy costs and is not incurred to meet peak reliability needs. This means that a significant portion of investment in generation, transmission and distribution plant (and the associated operation and maintenance expense) cannot be reasonably described as demand-related or driven by peak reliability needs. Any pricing structures that reflect the marginal costs of peak system capacity should be sized properly to reflect these distinctions. That includes

demand charges, if appropriate, as well as time-varying energy pricing.

It is fair to ask whether a properly sized "peak window" demand charge solves these issues. Although such a charge is superior to traditional demand charges for the pricing of shared capacity costs, peak window demand charges nonetheless retain many of the shortcomings of their traditional counterparts. Customers who have high usage at many times throughout the peak period should be charged more for capacity than customers who have a single high usage interval in that same window. Time-varying energy pricing provides superior incentives to optimize usage at all relevant times. Simple time-of-use rates are fairer and more efficient than peak window demand charges and can be made even more so by overlaying them with pricing that is responsive to critical peak conditions.

A few analysts and economists have identified several narrower applications where pricing structures akin to demand charges could be appropriate and reasonably efficient: (1) site infrastructure for individual customers, (2) risks related to customer variability at peak times and (3) timer peaks. While more research into these applications might be merited, demand-based pricing would only be a second-best approximation of a more efficient but potentially more administratively complex time- and location-based pricing system.

1. Introduction

Demand charges have existed almost since the beginning of the electric industry in the 1880s. They were originally called Hopkinson rates after John Hopkinson, a British engineer who described the concept in 1892. Hopkinson believed that costs of "plant and conductors"[1] — namely capacity costs — for an electric utility should be charged to customers based on the "greatest rate of supply the consumer will ever take."[2] Shortly thereafter, a meter was developed that could capture the highest kW power draw from the customer, defined over a period of an hour or half-hour, during an entire billing period (now typically a month). These rates became prevalent for industrial customers in the early 1900s.[3]

It did not take long, however, before economists called into question their putative cost-causation rationale. In 1941, future Nobel Prize winner in economics W. Arthur Lewis argued that the cost-causation case for demand charges was often based on "a simple confusion. … The maximum rate at which the individual consumer takes is irrelevant; what matters is how much he is taking at the time of the station's peak."[4] In 1970, prior to becoming chairman of the New York Public Service Commission, Cornell University professor Alfred E. Kahn wrote that demand charges are "basically illogical."[5] More recently, University of California professor and California Independent System Operator board member Severin Borenstein opined that "it is unclear why demand charges still exist."[6]

Electric utilities and some consultants still make broad arguments for demand charges that are, at their core, the same as those made more than a century ago. In 2016, the Edison Electric Institute (EEI) asserted that "demand charges provide accurate price signals" and "better collect capacity costs [than other kinds of prices]."[7] EEI made this

---

[1] Hopkinson, J. (1901). *Original papers: Vol. 1, Technical papers*, p. 257. Cambridge University Press.

[2] Hopkinson, 1901, p. 261.

[3] There was a debate within the electric utility industry about rate design in the 1890s. A time-of-use meter was invented nearly simultaneously with the demand meter, and some industry participants argued that time-of-use rates would be superior. See Hausman, W. J., & Neufeld, J. L. (1984). Time-of-day pricing in the U.S. electric power industry at the turn of the century. *The RAND Journal of Economics, 15*(1). This time-of-use meter disappeared from discussion relatively quickly, however, as an industry consensus formed around demand charges. Neufeld argues that demand charges were a part of utility strategy to discourage industrial customers from relying on distributed generation, known as "isolated plants" at the time. Neufeld, J. (1987, September). Price discrimination and the adoption of the electricity demand charge. *Journal of Economic History, 47*(3), 693-709. In addition, Samuel Insull, president of Chicago Edison (later Commonwealth Edison) and a major player in the industry, happened to own a part of the patent for the demand charge meter. See Yakubovich, V., Granovetter, M., & McGuire, P. (2005). Electric charges: The social construction of rate systems. *Theory and Society, 34,* 597-612.

[4] Lewis, W. A. (1941). The two-part tariff. *Economica, 8*(41), 252.

[5] Kahn, A. E. (1970). *The economics of regulation: Principles and institutions: Vol. 1, Economic principles*, p. 96. John Wiley & Sons.

[6] Borenstein, S. (2016). The economics of fixed price recovery. *The Electricity Journal, 29*(7), 10.

[7] Edison Electric Institute. (2016, February). *Primer on rate design for residential distributed generation,* p. 6. https://www.eei.org/issuesandpolicy/generation/NetMetering/Documents/2016%20Feb%20NARUC%20Primer%20on%20Rate%20Design.pdf

argument simultaneously for two different types of demand charges: (1) the traditional monthly noncoincident peak (NCP)[8] demand charge, based on an individual customer's NCP across an entire billing period, and (2) a peak window demand charge,[9] based on an individual customer NCP within a defined multihour interval, similar to the on-peak period for a time-of-use (TOU) rate.[10] In addition, there has been a push by EEI and many utilities to expand the application of demand charges beyond just the industrial and large commercial customer classes to small business and even residential customer classes.

Demand charges as we've known them in the United States should largely become a relic of the past. Current forms of demand charges, based on 15-minute, 30-minute or 60-minute individual customer peaks and often intended to recover the lion's share of capacity costs, are neither cost reflective nor efficient in general.[11] For much of the 20th century, traditional demand charges may have been a second-best alternative that worked reasonably well for high-load-factor industrial customers. Developments of the past several decades have, however, made even this application of demand charges archaic. Such charges do not reflect the cost drivers of the modern electric system, and typical sizing of these charges are larger than justified by proper economic analysis of the electric system. Peak window demand charges, while an improvement over their traditional counterpart, do not solve many of the core deficiencies of demand charges as an efficient pricing mechanism. Time-varying rates, including TOU rates and critical peak pricing, are more efficient than peak window demand charges.

If there is a role for demand charges in today's electric system, it is much narrower than the one it performs for industrial customers in many jurisdictions. Modern versions of

---

[8] A customer's noncoincident peak is its highest demand, in kilowatts, measured at the meter during the period in question. This customer demand can be measured based on different intervals, typically 15, 30 or 60 minutes. "Noncoincident" means that this demand does not necessarily occur at the time of a system peak.

[9] There is no standardized terminology for this type of demand charge where determination of the maximum demand for the billing period considers only a limited number of peak hours, similar to the peak period for a time-of-use rate. We find the "peak window demand charge" description more apt than the other alternatives.

[10] Less commonly, daily-as-used demand charges are part of the discussion, which we raise later in this paper. As the name implies, it is a demand charge for a customer's individual NCP in a given 24-hour period, sometimes limited to a peak window within that day and sometimes excluding weekends and holidays. This means that the ratchet feature of a daily-as-used demand charge is reset every day and not every billing period, as with other demand charges. In this paper, we do not focus on contract (ex ante) demand charges, although they share many features with these other alternatives.

[11] There are other issues at play in the debate around demand charges, particularly whether residential and small business customers can understand and manage these types of rates and the related potential for inequitable bill impacts. See Chernick, P., Colgan, J., Gilliam, R., Jester, D., & LeBel, M. (2016). *Charge without a cause? Assessing electric utility demand charges on small consumers.* (Electricity rate design review paper No. 1). https://votesolar.org/files/6414/6888/3283/Charge-Without-CauseFinal_71816.pdf; and Lazar J. (2015). Use great caution in design of residential demand charges. *Natural Gas & Electricity*. https://www.raponline.org/wp-content/uploads/2016/05/lazar-demandcharges-ngejournal-2015-dec.pdf. The question of understandability of demand charges by residential and small business customers is a longstanding one. D. J. Bolton notes that a 1948 report by a government commission in Great Britain rejected demand charges for residential customers on two bases: (1) understanding of the rate and (2) the potential reaction to an overload encouraging higher usage levels going forward. Bolton, D. J. (1951). *Electrical engineering economics: Vol. 2, Costs and tariffs in electricity supply*, p. 255. (2nd ed. rev.). Chapman & Hall. We do not delve into these issues at length in this paper.

these charges need to be more rigorously fashioned to achieve economic efficiency and advance the public good than they have been historically. We examine three more nuanced cases where demand charges have been identified as a potentially efficient pricing mechanism: (1) site infrastructure for individual customers, (2) risks related to customer variability at peak times and (3) timer peaks. In these situations, pricing structures with some similarities to demand charges may be appropriate. In each of these cases, demand-based pricing would only be a second-best approximation of a more efficient time- and location-based pricing system.

Unless we reexamine fundamental ratemaking practices critically in light of the modern electric system and new technologies, we will miss major opportunities to optimize system costs, ensure reliability and improve societal outcomes. While utilities and some consultants have been pushing for new applications for demand charges, regulators and utilities should be moving in the opposite direction by replacing demand charges for industrial customers with more accurate pricing mechanisms.

2. Historic Cost-Causation Argument for Demand Charges

A frequently used but inaccurate cost-causation argument for demand charges begins with the observation that several of the most important cost categories can be denoted in kilowatts (kW) or megawatts (MW).

- Generation capacity is denominated in kW or MW, reflecting the maximum instantaneous power output of a given unit.
- Transformers are rated in kilovolt-ampere (kVA) or megavolt-ampere (MVA), a unit of apparent power[12] closely related to kW or MW.
- Conductors are rated in amps for the level of current that they can handle. For a given voltage, this leads to a maximum kW or MW power flow for that conductor (power equals current times voltage).[13]

From these engineering descriptions, which are accurate but potentially misleading, some analysts conclude that, because generation and delivery capacity can be measured in units of power (kW or MW), their costs are demand-related. Making the leap to retail rate design then becomes easy: Capacity costs are rated in kW, so prices should be reflected

[12] Apparent power is the combination of active and reactive power in an alternating current circuit that needs to be supplied to serve load. This includes the power components that are needed to energize the circuit but don't transfer useful power to the load.

[13] For three-phase power, power is current times voltage times the square root of 3.

in kW.[14] This is the essence of the argument made by EEI, but it rests on several fallacies.

Some earlier writers, including W. Arthur Lewis, D. J. Bolton and James Bonbright, are open to demand charges to a certain extent but are quite candid about their limitations and significant downsides. Important factors in this more nuanced determination include:

- The diversity and coincidence factors[15] of any group of customers who might face a demand charge.
- The relative metering costs for flat kilowatt-hour (kWh) rates, demand rates and time-varying rates.
- The ability (or lack thereof) for customers to economically shift certain types of load.
- The broad similarity of capacity and fuel costs for many generation alternatives (typically thermal steam units) prior to 1960.

Lewis acknowledged the metering problem in his 1941 article, "The Two-Part Tariff." He stated that "the two-part tariff [a demand charge and an energy charge] is superior to having a single undifferentiated price which discourages off-peak consumption, *but inferior to charging different prices at different times*, though it may sometimes be more convenient than the latter if the measurement and timing of consumption are costly."[16] In the early and mid-20th century, only simple kWh metering was economic for small customers (that is, the system benefit from the response to time-differentiated pricing did not exceed the cost of the metering necessary to support it), while more sophisticated metering could be justified for industrial customers.

In 1951 Bolton noted, with some approval, that demand charges were much more common for industrial customers than residential.[17] He observed that residential customers' peaks are more random, that is to say more diverse (spread out in time) and less likely to be correlated with system peaks: "A metered demand system for such a [residential] consumer would mean making a high charge for payment at times when it was most unlikely to matter."[18] He opined that the load of many large industrial customers is not

[14] It is worth noting that these "kW" demand measurements are actually measured in units of kilowatt-hours per hour and simplified to be presented as measures of kW demand.

[15] Diversity of demand for a utility reflects the temporal differences in usage among customers. Peak coincident demand at any level of the system is less than the sum of customers' individual peaks because of these temporal differences. The calculated "diversity factor" provides a quantitative measure of these differences; conversely, a "coincidence factor" measures the extent to which these individual peaks do line up. These concepts are defined and discussed further in Section 3.1.

[16] Lewis, 1941, pp. 255-256 (emphasis added). Even in 1941, Lewis thought that it was no longer the case that demand metering would be cheaper than time-based metering, with one alternative being simple timers and another being "ripple control," where a utility sends a high frequency signal to flip an equipment switch.

[17] Bolton, 1951, p. 255. Bolton's proposed ideal "scientific tariff" features a TOU rate and no demand charges, where the on-peak price recovers demand-related costs. See Bolton, 1951, pp. 249-250.

[18] Bolton, 1951, p. 255.

particularly susceptible to shaping, because it is "motive power" (i.e., motors to run large equipment), and the electricity costs represent a small fraction of overall costs for these firms.[19] This type of industrial customer has strong incentives, given a set amount of productive capacity, to have the highest operating factor possible and thus a high load factor.[20] This industrial load pattern implies a significant likelihood that an individual customer's peak in a given month or year is closely linked to the customer's demand at the time of system peak.

Bonbright, writing originally in 1961, stated that traditional demand charges provide some benefits from "a tendency of existing customers to spread their loads over a longer period in order to minimize their demand charges, instead of bunching them during short period likely to coincide with the heavy loads of other customers."[21] Bonbright then went on to observe that electric rate design in those days "[was] far from ideal, and practical rate makers will do well to consider seriously its alleged infirmities viewed from the standpoint of its critics among the academic economists." He noted in particular that there was little sense in "the imposition of demand charges which penalize consumers for high individual demands even though these demands come at hours or seasons that fall well off the peak loads imposed on the system as a whole or even on any major part thereof."[22]

Up until 1960, most generation options, with the exception of hydroelectric power, had very similar cost characteristics. Steam generation was the predominant capacity type, and there were few differences in cost among coal, oil and natural gas units. Even fuel prices were broadly similar. In such a system, there is a better case that all capacity is similarly situated to serve peak reliability needs and thus can be considered demand-related. As discussed later, this issue goes to the sizing of any demand charges if they can be shown to be a reasonable solution (in limited circumstances at best).

This combination of factors — (1) an industrial customer base with a relatively small number of customers, most of whom had high load factors, high peak-coincidence factors and high levels of consumption and (2) a large number of residential customers with lower coincidence factors and relatively low consumption per customer — provided a rough rationale for the rate designs that prevailed throughout most of the United States in the 20th century and are now lingering into the 21st. In pricing, this typically manifested itself in significant demand charges for large industrial classes to recover nearly all capacity costs and in fully volumetric energy rates for residential and small business customers.

---

[19] Bolton, 1951, p. 238.

[20] Load factor is the ratio of an end user's actual energy usage in a period to its maximum potential usage in that period. It is calculated as follows: kWh/(peak demand x total hours), within the specified period.

[21] Bonbright, J. (1961). *Principles of public utility rates*, p. 311. Columbia University Press. https://www.raponline.org/knowledge-center/principles-of-public-utility-rates/

[22] Bonbright, 1961, p. 316.

In this historical context, this could be relatively fair and efficient for a narrow slice of customers that meet the relevant description. To the extent that other customers that could share capacity (e.g., churches and schools; offices and movie theaters) were faced with demand charges, these customers were treated unfairly and often paid significantly more costs than they caused.

3. Why Demand Charges Are Inefficient

Some of these arguments for demand charges held sway in the past, even though the better case for time-varying energy charges was well understood. Today, the features of the modern electric system undermine even the more nuanced historic case for demand charges altogether. This is true for large industrial customers as well as for residential and small business customers.

The original advocates of demand charges often focused on what they thought was a fair and efficient division of historic accounting costs. Modern economists, even those who still advocate for demand charges, recognize that this older perspective is in error and argue (correctly) that rate structures should be designed to efficiently optimize *future* costs.[23] This perspective leads one to the conclusion that rates should be reflective of forward-looking marginal costs. In utility regulation, this concept is translated into different operative regulatory language in different jurisdictions, calling variously for rates that discourage wasteful usage, reflect actual costs or ensure the causer pays those costs. But in each case, the underlying microeconomic principle is the same: Rate design should ensure that the actions customers take to minimize their own bills are consistent with the actions they would take to minimize system costs. The nitty-gritty of designing rates in this framework is how to fairly and efficiently reflect marginal costs in prices. The best way to conceptualize this is to examine how the customer responds to a given rate design — both its form and its magnitude. An efficient rate design will lead to customer behavior that optimizes system costs.

The marginal consumption incentives for customers in any system of time-varying rates are fairly straightforward: (1) discourage usage in periods of relatively high rates and (2) encourage usage in periods of relatively low rates. Prices that achieve these outcomes are charged in a way that is both (1) consistent (all kWh at a given time or system condition are treated the same) and (2) symmetric: If an increase in consumption causes a bill to rise by $10, then the same sized decrease causes a bill to decline by $10.

The incentives presented by a typical demand charge structure are somewhat more

[23] See, for example, Boiteux, M. (1960). Peak-load pricing. *The Journal of Business, 33*(2), 157-179. (H. W. Izzzard, Trans.); Kahn, 1970; and Crew, M. A., & Kleindorfer, P. R. (1979). *Public utility economics.* St. Martin's Press.

complex.[24] If a customer is perfectly flexible (indifferent as to when they take electricity from the grid) and has perfect foresight, a demand charge would clearly incentivize a 100% load factor within the relevant time frame (e.g., each month). Of course, such customers do not exist in the real world,[25] although there are some customers that come close to having 100% load factors because of the nature of their operations: 24-hour supermarkets, data centers and certain types of factories.

Because customers do not have perfect foresight and infinite flexibility, it is only possible to talk about the incentives created by a demand charge at a certain level of generality. The most obvious features of a demand charge are that it directly (1) discourages higher individual customer NCP demand and (2) encourages levelization of load within the relevant time period. The related key feature of all types of demand charges is that they act as a ratchet, even if the ratchet is not applied across multiple billing periods.[26] Once a certain level of demand has been reached, customers then face a *lower* marginal cost for the remainder of the period to which the demand charge applies, as long as they have a power draw between zero and their previous individual demand peaks.

When the demand charge impacts a particular consumption decision, it can be quite punitive — imagine paying $5 to $10 to make toast for a family, which is exactly what can happen with a poorly designed residential demand charge.[27] This shows up as a high marginal cost for a subset of hours and consumption decisions. But otherwise, if a particular consumption decision does not pose a substantial risk of setting the demand charge, then consumption becomes cheaper — defined solely by the other charges without any demand charge implications. This means that optimal customer decision-making under a demand charge is quite complex and depends on the level of foresight and the value of consumption across all of the relevant time periods. Of course, most customer decision-making will not necessarily be optimal but rather based on rules of thumb, particularly for residential and smaller commercial customers.

---

[24] Sandford Berg and Andreas Savvides did some theoretical work that incorporated the granular incentives of a demand charge into a traditional economic model of consumption. See Berg, S. V., & Savvides, A. (1983, October). The theory of maximum kW demand charges for electricity. *Energy Economics (5)*4, 258-66. However, this was a two-period model with numerous simplifying assumptions. Such a simplified theoretical model does illuminate certain features of a demand charge, but the authors note numerous areas for further work. To our knowledge, this line of theoretical research has not been pursued.

[25] This is true in particular because customer "utility" from electricity is not solely about the amount of consumption. Customers also enjoy significant convenience benefits for certain usage timing, again assuming that on-site storage and energy management are not cheap and convenient enough to smooth these features out.

[26] Some rates that do not meet this criterion are occasionally described as demand charges, such as annual system coincident peak capacity charges. These types of charges may, however, be better thought of as a type of time-varying rate or perhaps in a third category of their own.

[27] A toaster is approximately 1 kW demand; see Home Energy Saver & Score: Engineering Documentation. (n.d.). *Default energy consumption of MELs*. http://hes-documentation.lbl.gov/calculation-methodology/calculation-of-energy-consumption/major-appliances/miscellaneous-equipment-energy-consumption/default-energy-consumption-of-mels. If a customer uses it for 15 minutes straight at the time of the customer's individual peak, the monthly demand billing determinant increases by 1 kW with a corresponding bill increase.

As the analysis in the subsections that follow shows, demand charges — whether of the traditional monthly variety or the peak window variety — are inefficient and inequitable for the pricing of shared system costs, as is the continued reliance on them. There are three interrelated reasons for this:

1. Traditional monthly demand charges provide an inaccurate price signal that is unrelated to high-cost periods for nearly all customers and which leads to inefficient customer efforts and investments in response to its incentives. The changes in the electric system due to dramatic increase in wind and solar generation mean that, from a system perspective, very high industrial load factors are not necessarily optimal.
2. Even in cases where a traditional demand charge could be justified, the sizing of demand charges to recover nearly all generation and delivery capacity costs reflects an outdated perspective of the engineering and economics of the electric system. Modern cost allocation and rate design must reflect the trade-offs between different types of expenses and investments. Much capacity investment is designed to reduce energy costs and line losses and should be charged on that basis.
3. Although a reasonably sized peak window demand charge is superior to a traditional monthly demand charge, time-of-use and other kinds of time-varying rates remain more efficient and equitable. These time-varying rate options are enabled by the dramatic decrease in the cost of sophisticated metering over the past two decades.

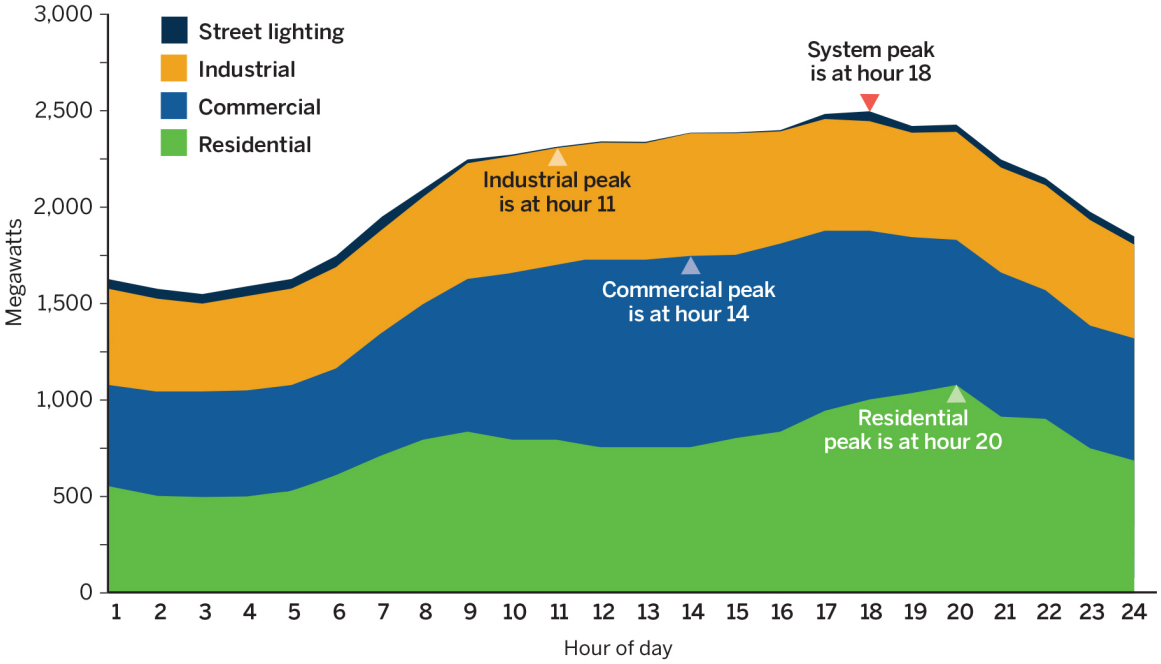
3.1 Individual Peaks Are Not the Same as System Peaks

Virtually all of the electric system consists of capacity that is shared among customers. With the exception of facilities that serve one or a very few customers, each component of the system is sized to meet an expected peak coincident demand of the customers it serves. The costs incurred to meet peak coincident demand, both short-run variable costs and capacity investment, are a significant portion of overall system costs. As a matter of economic efficiency, it is crucial that prices reflect the marginal costs of meeting the coincident system peak. Peak coincident demand is not simply the sum of the customers' individual peak demands but is rather something less, often significantly so. This phenomenon is known as *diversity* of demand and reflects the temporal differences of usage across the relevant customer base.

Customer loads are diversified at every level of the utility system. At the system level, the peak is determined by that combination of customer class loads that produces the highest instantaneous demand. That system peak might, or might not, coincide with the peak demand of any one customer class, and that system is likely interconnected to other systems with slightly different loads, through a shared transmission network. Figure 1 shows illustrative customer class loads on a system peak day. Each of the customer classes has a highest load hour at a different time: hour 11 for industrial, hour 14 for commercial

and hour 20 for residential. The load for the lighting class is roughly the same across many different hours when the sun is down. The overall peak is at hour 18, which is different than any of the class peaks.

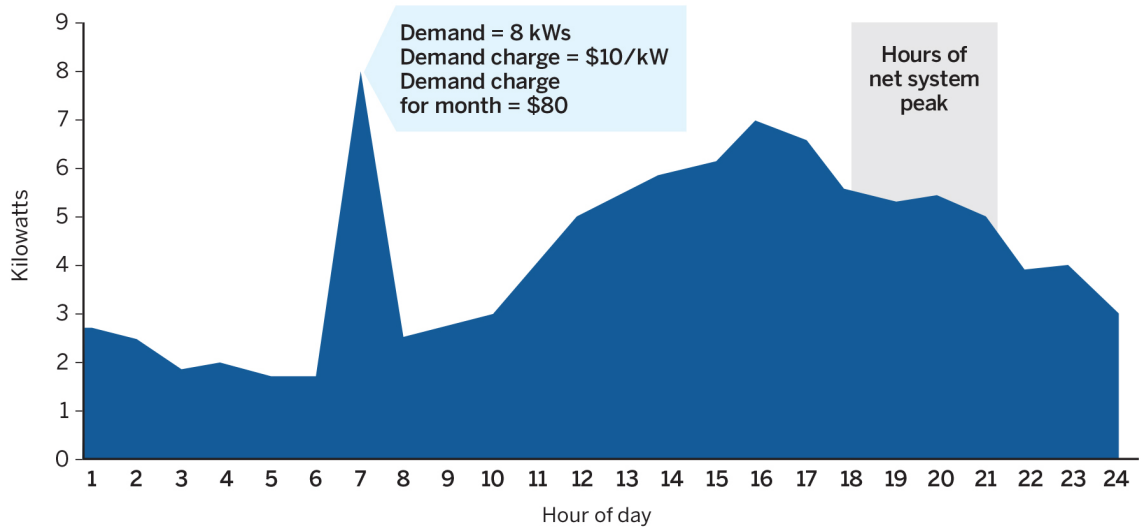
Figure 1. Diversity at the customer class level



Diversity can be quantified as the ratio of the sum of the subgroup peaks to the relevant coincident peak — the diversity factor. In this illustrative example, the diversity factor of the customer classes is 1.1. Diversity factors cannot go below 1 because in the extreme case where all subgroups peak at the same time, the sum of the subgroups equals the overall coincident peak. As long as customers peak at different times, diversity factors are higher as you consider smaller subgroups. Load diversity across individual customers is even greater than across customer classes.

Traditional monthly demand charges impose a rate on each customer that is independent of the system peak, as illustrated in Figure 2 on the next page. These demand charges provide little, if any, incentive to minimize a customer's contribution to system peak, unless a strong correlation exists between the customer's peak and the system's, a circumstance known as a high coincidence factor. In this illustrative example, a residential customer has an electric water heater that runs for nearly a full hour in the morning and a substantial cooling load in the afternoon.

Figure 2. Illustrative monthly noncoincident peak demand charge for an individual residential customer



Demand charges encourage customers to flatten their own load curves relative to their individual maximum usage but do not necessarily encourage them to consume energy in ways that optimize system costs. If we assume that Figure 2 shows customer usage before a traditional monthly demand charge is imposed, we could expect significant changes in usage after application of this charge. It would be reasonable to expect this customer to attempt to reduce the 8 kW demand reached at 7 a.m. In the case of an electric water heater, the individuals living in the house could change their behavior or adjust the settings on the water heater. If the customer could reduce that morning peak, then there would be some incentive to reduce the afternoon peak caused predominantly by cooling load. In this case, the customer would benefit by moving some portion of that load away from hour 16 to other hours, including possibly during the system peak from hours 18 to 21. Furthermore, this customer could increase overall kWh consumption since the marginal cost would be lower at times (often including the system peak) when there is little risk of triggering a higher demand charge.

More generally, a flat individual customer load shape may not, in fact, be what is best for the system and is in fact worse than a low load factor with predominantly off-peak usage. The clearest illustration of this is street lighting load, which, for most systems, falls entirely outside the system peak hours and has a roughly 50% load factor. If we designed and sized a demand charge for street lighting on the same basis as a typical demand charge for industrial customers, it would force this low-cost off-peak load to pay as much for system capacity as an industrial customer using the same amount of power during the peak periods. This is virtually never done, however, and street lighting is treated as a separate rate class without any demand charges.

D. J. Bolton summarized the basic problem facing utilities and regulators in the middle of the 20th century:

> The aim should always be the improvement of the *system* load factor, and the only justification for an elaborate tariff is that it shall contribute directly to this end. … If these costs are passed on to the consumer as they stand, in the form of a two-part [maximum demand] tariff, the fixed charge will be levied on the consumer's individual [maximum demand] instead of his effective demand on the system. The consequence will be that low-load-factor consumers will be overcharged (since they are given insufficient credit for their greater diversity) whilst the high-load-factor consumers are under-charged.
>
> The weakness of such a tariff when applied to the small individual consumer is that it treats load factor as a variable and diversity factor as a constant. … But, in practice, diversity factors vary from consumer to consumer almost as much as load factors, and moreover, in the opposite direction.[28]

In other words, diverse customers can efficiently *share* capacity, and rate design should recognize this fact. As Bolton mentioned, it is often the case that small users have *lower* load factors but more diversity and thus less impact on peak. This is still true today because many small residential users have lower levels of heating and cooling usage (smaller residences) and often have similar appliances (microwaves, toasters, dishwashers and dryers) that are used more sporadically than larger residential customers. This means that the load factor for each individual appliance is lower, but the power characteristics are similar for each usage of an appliance.

As described in Section 2, demand charges may present a rough price signal to control peak system demand for customers with a high system-peak coincidence factor. In that case, controlling a customer's individual peak does systematically reduce the overall coincident peak. One case where this could be true historically is large industrial customer classes, where individual customer usage is driven by large equipment that is constantly used throughout every working day of the year. Even for this type of customer, however, there remains the question of whether load can be shifted from peak hours to off-peak hours. A critical peak energy price would produce a superior price signal, to actively reduce usage at critical peak hours, rather than maintain steady usage at those hours if such a shift is possible. Indeed, industrial customers in Texas, faced with significant, narrowly focused transmission charges based on four coincident peak hours, use specialized consultants to help them identify, in advance, the hours to which those

[28] Bolton, 1951, p. 107-108 (emphasis in original). Bolton was writing at a time when, in operations, customer demand was taken largely as a given and much of the resource mix was dispatchable thermal generation. In those circumstances, improvement of system load factor would, all else again being equal including overall kWh consumption, lead to a reduction in total system costs.

charges will be applied and reduce usage sharply in those hours.[29]

For a diverse customer class, however, the share of customers who face this demand charge price signal at system peak times is random and inconsistent. In almost any hour, whether near system peak or the lowest-load hours of the year, some customers will face the demand charge price signal. Also, a substantial number and, at times, a majority of customers (e.g., those customers who have already hit their peaks in the billing period) face a lower marginal cost at system peak times. While this is very blunt and inaccurate, it could be a sharper price signal than a traditional flat kWh rate in some circumstances, although a customer's likelihood of facing those circumstances would vary randomly. In contrast, a well-designed TOU rate provides the broadly correct incentive for all marginal consumption choices by all customers, sending a consistent price signal for on-peak and off-peak periods; a critical peak pricing rate can be even more precise, focusing on specific hours when the electric system is under stress.[30]

The undesirable effects of demand charges are made worse by ratchets across billing periods — the mechanism by which a maximum demand in one period becomes the basis for minimum billed demand in subsequent periods. For example, billing demand may be the greater of this month's noncoincident maximum load and 80% of maximum in the previous 12 months. Once a maximum demand is hit, the customer has little incentive to reduce demand in the following periods. Unless individual customer peak is closely linked to system peak, there remains little incentive to minimize usage at a time of system peak.

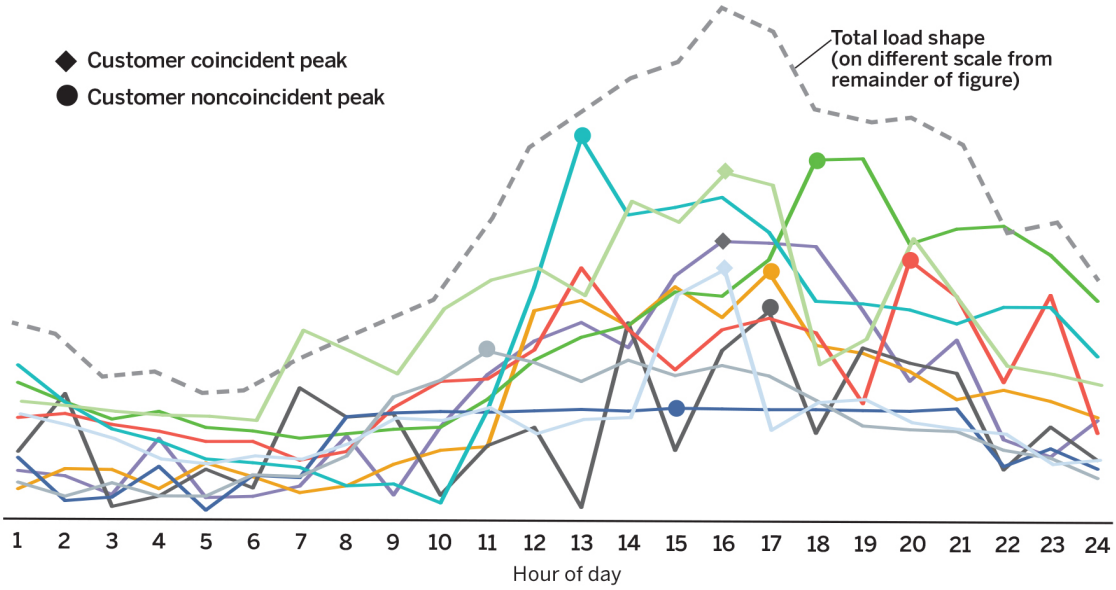
It is only when one gets close to the end user that the components of the system — the final line transformers, secondary distribution lines and service lines — are sized to meet a very localized demand that can be directly attributed to a small number of customers. Even at this level, there can be significant diversity among customers sharing a single transformer.

[29] Zarnikau, J., & Thal, D. (2013, September). The response of large industrial energy consumers to four coincident peak (4CP) transmission charges in the Texas (ERCOT) market. *Utilities Policy, 26*, 1-6.

[30] To be more precise, we should say a "relevant component of the system" since different components of the system may hit peaks at different times. It's not unusual to see a systemwide peak occur at a particular hour on a particular day, but for individual elements of the subtransmission and distribution systems to hit peaks at other times. Expressing these peaks in prices and capturing each user's causal relationship to them is a challenge of time-varying rate design and, to the extent that this reflects different peaks in different areas of the system, may require locational distinctions as well. Precision is valuable, but complexity may produce inferior customer response.

Figure 3 shows actual data from a confidential load research sample on a summer peak day for 10 residential customers who share a line transformer. The total load shape is on a different scale than the individual customer loads.

Figure 3. Summer peak day load from 10 residential customers on one line transformer

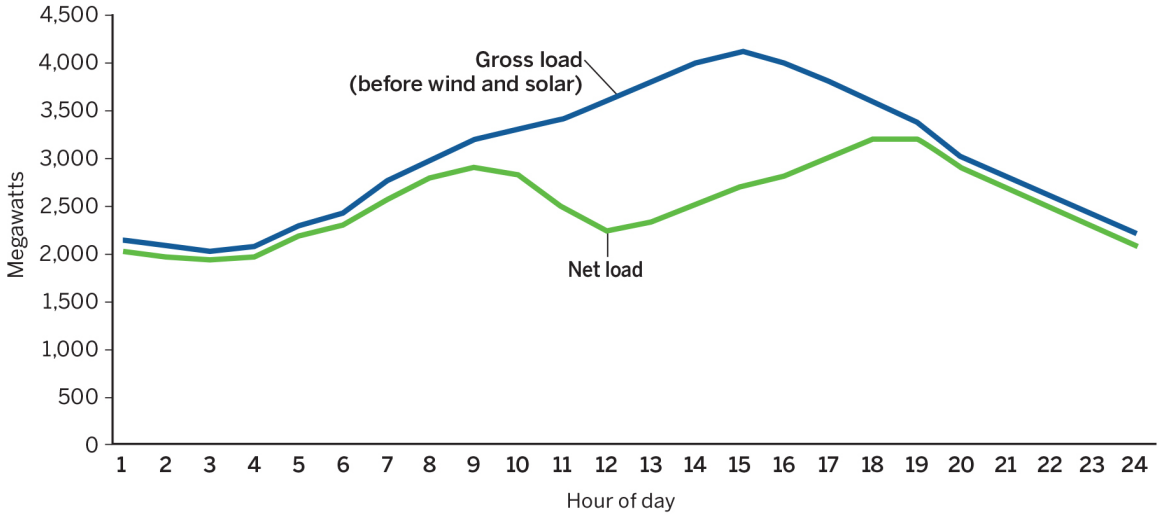


Source: Confidential load research sample

This demonstrates how diversity determines the need for the sizing of system elements. Only three of the 10 customers peak at the same time as the 4 p.m. coincident peak for the group, and the coincident peak is only 86% of the sum of the individual peaks on this day, which translates into a diversity factor of 1.16. This is just the variation on a particular high-load day. Although not shown in this figure, this coincident peak is only 64% of the sum of the annual NCPs for the individual customers, which translates into a diversity factor of 1.56.

At least two features of the modern electric system are changing the traditional argument that high-load-factor industrial customers should be subject to demand charges. First, the timing of traditional peaks and valleys, and by extension their effect on both short-run variable costs and longer-term capacity needs, is changing due to the increased prevalence of variable renewable resources. In regions where solar generation has increased rapidly, the “duck” curve is now a familiar phenomenon, as shown in Figure 4. Second, relatively low-cost on-site energy storage means that *all customers* have the potential for economically shiftable load and can respond to time-based price signals.

Figure 4. Illustrative net load curve



In such a situation, the benefits of shifting energy intensive industrial load from early evening to midday could be quite large. But this could mean *an increase* in the customer load factor, which is substantially discouraged by demand charges.

3.2 A Significant Portion of Capacity Investment Is Not Demand-Related

Traditional cost allocation terminology makes a distinction among demand-related, energy-related and customer-related costs. This terminology may obscure more than it illuminates. In particular, the term "demand-related" is often used to imply that demand charges are a proper pricing method for recovering costs so designated. Moreover, "demand-related" has typically referred to system peak demand and not individual customer peaks.[31] Other terminology, such as "peak-related," is more descriptive of the concept and avoids confusion with the use of "demand" in other contexts (such as "demand for energy").

[31] See Bolton, 1951, p. 132 (describing demand-related costs as "a cost proportional to system demand") and pp. 143-144 (describing how to spread costs across a wide number of potential system peak hours). In rate design, these same costs might be recovered through demand charges for certain customer classes. When determining the rate in dollars per kW, the total costs are then divided over the larger denominator of individual NCP demand, without accounting for load diversity within the class. This reduces the dollars per kW as charged to each customer from the dollars per kW used to assign costs to each class. This reduction is labeled differently in different jurisdictions, such as an "effective demand factor." However, this reduction is passed through to all customers and does not correct for differences in the timing of individual customer peaks. Customers who have demand highest at peak times receive a discount, and those who have demand highest at other times are overcharged.

Advocates for demand charges sometimes assert that most or all capacity costs are demand-related, which maximizes the size of the demand charge (if one is at all justified).[32] This leads to the large magnitude of the demand charges for industrial customer classes in many states. However, significant portions of capacity costs are not demand-related but are in fact incurred to meet energy needs. Investments in generation, transmission and distribution in the modern electric system may serve either of the two primary objectives of system planners, but the degree to which demand plays a role in each objective is different. These two goals are: (1) ensuring reliability (in both operational and investment time frames) and (2) meeting year-round system load at least cost. In many respects, reliability concerns arise predominantly at peak system hours.[33] Meeting system load at least cost, by its very nature, must consider usage patterns across every hour of the year. To meet these two objectives, system planning, investment and operation must jointly consider not only the engineering and physics of the electric system but also the economics of the relevant choices. We see this tension in the evolving landscape of capacity resources.

With respect to generation, most capacity costs may have been demand-related prior to the invention of the modern combustion turbine in the 1960s. In an electric system dominated by largely homogenous steam generation capacity, a MW of capacity built for peak demand could be used equivalently year-round.[34] In such a situation, generation capacity costs could be allocated and charged predominantly at peak times.

The existence of multiple different types of generation capacity, storage and demand response changes this analysis significantly.[35] Aggregate supply (generation, storage and demand response) must be sufficient for systemwide coincident peaks, as well as contingencies across many other hours of the year, such as when outages (unforced and even planned, such as nuclear refueling) combine with other circumstances (e.g., unusual weather) to push demand up against the limit of available resources.

---

[32] See Faruqui, A., & Davis, W. (2016, July). Curating the future of residential rate design. *Electricity Daily*, 23. http://files.brattle.com/files/7137_curating_the_future_of_rate_design_for_residential_customers.pdf. The authors state that "a large share of a utility's costs are actually driven by investment in infrastructure, such as generation capacity and transmission and distribution (T&D) networks. These costs are not directly related to the amount of energy that is consumed; they are, instead, driven by various measures of maximum electricity demand." See also the description of an idealized rate design that "recover[s] capacity costs through demand charges" in Faruqui, A. (2019, June 1). 2040: A pricing odyssey. *Public Utilities Fortnightly*, p. 56.

[33] Reliability can be thought of as having two dimensions, in terms of both system security and resource adequacy. The former refers to operational time frames, being assured that the system has sufficient resources to meet demand in real time. The latter refers to investment time frames, being assured that the system will continue to deploy needed capacity to reliably serve load over the longer term. Both kinds of reliability are relevant to this discussion.

[34] Even this historic scenario is a substantial oversimplification due to significant level of hydro generation in many areas.

[35] Bonbright recognized this briefly in a footnote; see Bonbright, 1961, 354, fn 15. By 1970, this was a better understood and less theoretical concept so that Kahn spent multiple pages discussing it; see Kahn, 1970, pp. 97-98. M. A. Crew and P. R. Kleindorfer formalized mathematical models of optimal pricing with multiple different types of generation capacity; see Crew & Kleindorfer, 1979.

The optimal mix of resource types depends on the broader load patterns. Different generation technologies have different capabilities and different cost characteristics and should not be blindly lumped together as "capacity" for cost allocation and rate design purposes. The kind of capacity that one would build to meet short-term coincident peak needs, as well as reserves on short notice throughout the year, is much different than the kind of capacity that one would build to generate year-round. Indeed, for very infrequent needs, demand response (paying customers to curtail usage for a short period) is proving much cheaper than building *any* kind of generation resource that is seldom used. In order to be economic, capacity that serves only short-term needs must have low upfront investment costs, such as combustion turbines or demand response, but can have higher short-term variable costs when it is used. The combustion turbine is cheap to build but relatively inefficient and expensive to run. In contrast, a larger investment can only be justified by lower expected short-run variable generation costs and a higher expected capacity factor. As a result, this high-upfront-cost capacity lowers the total cost of both meeting peak demand and serving energy needs over the planning horizon.

So there is a trade-off between capacity costs and energy costs. Put simply, not all capacity costs are incurred to meet peak demand. As a result, capacity costs for generation should either be split into the traditional demand-related and energy-related categories, or else those categories should be updated into a more modern time-based classification framework.[36] Under any reasonable version of the demand-related classification, it is important to recognize that the capacity costs placed here are to serve relatively short-period peak reliability needs.

Even the appropriate short-period peak reliability capacity costs should be charged on a broader basis than the absolute peak hour of the year for several reasons. One is that, while planners and operators generally have a good idea of when a system peak is likely to occur, they by no means know for sure. Consequently, there is a reliability value to capacity in many hours that should be reflected in prices.[37] A second is that the actual peak can be influenced by pricing structures. For example, if a system peak could be reliably predicted for the 5 p.m. hour on a given day, charging a higher price at that single hour could just push that same peak to 4 p.m. without a meaningful reduction. This is the "whack-a-mole" problem. Taking both of these issues into account, some writers have referred to the relevant set of peak hours as the "potential peak" period.[38] This is a major consideration in the determination of on-peak hours for a TOU rate or a peak

[36] See Lazar, J., Chernick, P., Marcus, W., & LeBel, M. (Ed.) (2020). *Electric cost allocation for a new era: A manual.* Regulatory Assistance Project. https://www.raponline.org/knowledge-center/electric-cost-allocation-new-era/

[37] The operating reserves demand curve mechanism in the ERCOT wholesale market is one means of establishing that value across the entire year. In many areas, the loss of load probability is relatively high for only 50-100 hours per year, which is the typical design criteria for critical peak pricing and demand response programs.

[38] Bolton, 1951, p. 143.

window demand charge. A related challenge is that different elements of the system (e.g., generation, transmission and distribution) may peak at different times, which should be accounted for to the extent possible.

Generation capacity also has some reliability value in off-peak hours. Generation reliability issues may come primarily at peak times but certainly not exclusively. This can be because of generator outages (both planned and unplanned), unusual weather, transmission outages, other operating constraints or a combination of the above. D. J. Bolton commented in 1951 that there had been several times that load needed to be shed in off-peak seasons because of generator maintenance, which was "a definite indication of demand-related expenses on account of generating plant" even in off-peak seasons.[39] A loss-of-energy-expectation study calculates the year-round generation reliability risks and is one of the best ways to allocate demand-related generation capacity costs (but not energy-related generation capacity costs) over the entire year.[40] A probability-of-dispatch method, alternatively, assigns the total costs of generation resources to the hours in which each resource provides service.

Many of these same considerations apply to the transmission and distribution system, and an analyst should look to the underlying purposes and benefits of system investments to allocate and charge them properly. Several different kinds of transmission capacity are intended to deliver energy and are not designed primarily to meet reliability needs. The transmission segment that connects a generating unit to the broader transmission network can be properly thought of as a generation-related cost and charged on the same basis as the underlying generator. In many situations, long transmission lines are needed to connect low-cost generation resources, such as remote hydroelectric facilities or mine-mouth coal plants, to the network. These long lines are built to facilitate access to cheap energy and should be classified on that basis. Last, transmission lines built to facilitate exchanges between load zones are not necessarily most highly used at peak times but are used to optimize dispatch and trade energy across many hours of the year.

Other parts of the transmission and distribution network do need to be sized to meet peak demand and other reliability contingencies. But there are several different engineering options for transmission and distribution networks that have implications with respect to line losses, another clear energy-related benefit.[41] There are generally two types of losses incurred across the transmission and distribution system: no-load losses and load losses. No-load losses are incurred primarily to energize transformers (both station transformers

[39] Bolton, 1951, p. 143.

[40] Lazar et al., 2020, p. 132.

[41] See generally Lazar, J., & Baldwin, X. (2011). *Valuing the contribution of energy efficiency to avoided marginal line losses and reserve requirements.* Regulatory Assistance Project. https://www.raponline.org/knowledge-center/valuing-the-contribution-of-energy-efficiency-to-avoided-marginal-line-losses-and-reserve-requirements/

and line transformers). Smaller transformers consume less energy in this respect, but overloaded transformers incur high load-related losses, so optimal transformer sizing saves energy.

The system planning considerations for load losses, also known as resistive losses, are more complex. These losses occur as electrical current flows through each element of the system. These losses manifest themselves in the form of heat and reduce the amount of useful power that can supply customer loads. This relationship is represented by the formula:

Load losses (in kW) = I^2 x R

Where I = current (in amps) and R = resistance (in ohms)

Load losses can be decreased by reducing the resistance or reducing the current. Installing conductors with thicker metal wires is a simple way to reduce resistance, but these larger conductors are more expensive. Investments that reduce the current can, however, be much more effective because losses go up with the square of current. Any investment that reduces the current by 50% will reduce load losses by 75%, and any investment that reduces the current by 90% will reduce load losses by 99%. Since the current required to supply load is highest during peak demand periods, system losses are greatest during peak demand periods. There are several different types of capacity investments that reduce current substantially:

- **Higher voltage lines:** There is a direct relationship between the voltage of a line, the current passing through the line and the power delivered.
    - Current (in amps) = power (in kW)/voltage (in volts)
    - As a result, increasing the voltage by a factor of 10 reduces the current by 90%, which in turn reduces load losses by 99%.
- **Siting substations closer to loads:** By siting substations closer to loads, one can reduce the losses incurred by having conductors at lower voltages supply loads across long distances — the latter condition resulting in higher currents and relatively higher losses.
- **Converting single-phase distribution lines to three-phase power:** Three-phase power requires one additional conductor and additional space for the arrangement of the lines. For three phase lines, current = power/(voltage x √3). At the same voltage, current drops by 42.3% and load losses are reduced by two-thirds.
- **Distribution level control of voltage and reactive power:** Capacitor banks, smart PV inverters, voltage regulators and other more distributed assets across the system can compensate for voltage and reactive power needs at a local level that would

otherwise need to be met through the supply of upstream resources delivered through the grid — the latter condition resulting in higher currents and greater incurred losses.

- **Optimizing the location and size of line transformers:** Siting transformers closer to customers allows for shorter secondary lines that have low voltage and thus higher losses per foot. For some areas, this may require additional transformers, which comes at a cost. Smaller transformers also have lower no-load losses. Unfortunately, smaller transformers have lower rated capacities and thus higher load losses for a given level of current. Conversely, larger transformers have higher no-load losses but lower load losses. These complex economics should be analyzed to account for trade-offs between capital costs and energy losses. Modern advanced metering infrastructure (AMI) systems provide the ability to prepare heat maps on each transformer, enabling optimal sizing to minimize costs and losses.[42]

All of these factors should be accounted for in both cost allocation and rate design. Energy-related benefits from transmission and distribution capital investments are quite extensive. In a relevant sense, nearly all transmission lines are built with a substantial purpose of minimizing line losses for the delivery of large volumes of energy. Choice of the voltage level for a transmission line, either for a new line or upgrading an old line, involves higher capital costs for higher voltages with the counteracting benefit of lower losses. These costs are energy-related costs, not capacity-related costs. Furthermore, many of these energy benefits from investments to minimize line losses are not static over the course of the year. They increase dramatically at times of system peak because current delivered over the system is much higher, and marginal system losses at the time of peak can be 15-20% in many utility systems.[43] In addition, these benefits can be compounding because they are not limited to fuel costs or wholesale purchases. A more efficient transmission and distribution system can lower generation capacity requirements as well, including reserves.

All of these economic and engineering phenomena should be properly reflected in any analyses of cost causation. More specifically, these distinctions must be passed into rate design or else it gives rise to opportunities for customers to take inappropriate advantage by gaming the rates, with bill savings that far exceed any long-term reduction in system costs. The experience of the British Central Electricity Generating Board, a wholesale provider, provides a stark example of this in the late 1960s. The central board charged the regional boards for generation capacity costs based solely on a narrow peak window. In response, the regional area boards built their own combustion turbines at significantly lower cost to generate during these peak hours. This forced the central board to change its

[42] See Lazar, J. (2018, October 18). *Smart grid and community benefits — with no rate increase? How Burbank made it happen.* Regulatory Assistance Project. https://www.raponline.org/blog/smart-grid-and-community-benefits-with-no-rate-increase-how-burbank-made-it-happen/

[43] Lazar & Baldwin, 2011, p. 4.

wholesale rates, charging for only marginal capacity costs in a short peak and charging for the bulk of capacity costs in a broader period.[44] The key insight in this scenario is that demand-related costs charged to peak times should only reflect the marginal costs of relatively cheap generation, storage or demand response capacity costs incurred for short-period peak reliability purposes.

Modern examples of this pricing problem can be found in the current practices of several independent system operators and generation and transmission suppliers. For example, ERCOT currently charges on the basis of the highest hour in each of the four summer months for recovery of embedded transmission system costs to distribution service providers. This type of pricing mechanism is inappropriate for transmission costs and furthermore distorts the operation of the wholesale energy markets by over-incentivizing a wide range of customer actions.[45] Similarly, many electric cooperatives, charged by their generation and transmission suppliers on the basis of NCP demand imposed on the wholesale supplier, have installed water heater control systems to mitigate this demand at much lower cost than the avoided demand charges. Since the generation and transmission demand charges include the cost of baseload units and transmission, they greatly overstate the value of localized NCP load reductions. While these are wholesale examples, the same economic proposition also extends to retail rates.

3.3 Time-Varying Energy Rates Are More Efficient Than Peak Window Demand Charges

Once one acknowledges the time-dependent nature of cost in the generation and delivery of electricity to end users on a shared system, one must necessarily acknowledge the superiority (as matters of economic efficiency and fairness) of prices that reveal to those end users that temporal variability in cost to those that do not. The question, then, is simple: What should those prices look like? In some sense, a peak window demand charge does recognize this time dependency. However, a comparison of the incentives presented by time-varying demand charges and time-varying kWh charges reveals why time-varying kWh charges are the better approach.

There are several types of time-varying energy rates to be considered today.[46] Key design choices for these rates include the number of time periods, whether the price for each time period is set long in advance or can itself vary based on system conditions and market

[44] Kahn, 1970, pp. 97-98.

[45] See Hogan, W., & Pope, S. (2017, May). *Priorities for the evolution of an energy-only electricity market design in ERCOT*, pp. 69-79. Harvard University and FTI Consulting. https://hepg.hks.harvard.edu/publications/priorities-evolution-energy-only-electricity-market-design-ercot-0. We do not endorse the proposed solution of Hogan and Pope but agree with the transmission pricing problem that they describe.

[46] From this definition we exclude seasonal rates and kWh prices that vary from billing period to billing period. These kinds of rates can also reflect the cost-causation basis of rates but provide little or no incentive to manage usage within a billing period.

outcomes, and the actual prices for each time period.[47] The simplest is known as a time-of-use or time-of-day rate, which utilizes a small number of preset time periods and prices within each billing period. The most sophisticated time-varying rates are typically described as real-time prices, which are updated at short, regular intervals (e.g., hourly) based on prices in wholesale energy markets.

There are also options that combined preset time periods with pricing that varies based on system conditions in a predictable manner. With critical peak pricing, or the related peak-time rebate alternative, higher prices for times when the grid is stressed are set well in advance, but the days (and perhaps the hours) where these higher prices apply are actively chosen in response to system conditions. Variable peak pricing, as currently offered by Oklahoma Gas & Electric,[48] adds another layer of price differentiation by allowing more preset options for the on-peak price period. The on-peak price depends on market conditions: low, standard, high and critical. This choice between four different alternative on-peak prices allows for a higher level of precision in marginal incentives. All of these variations share a common goal — to improve the load shape for a utility by decreasing peak period load and shifting some of that to off-peak periods.

In this context, it is most natural to compare peak window demand charges with simple TOU rates because many of the key parameters can be kept constant. For both of these options, the peak time periods and the prices charged are set well in advance and can be set to recover the same categories of costs. Holding those two variables constant, peak window demand charges are inferior to time-varying kWh charges in that same peak window, as a general method for charging peak capacity costs, for two related reasons:

1. The inefficiency of the ratchet that all demand charges impose, which incorrectly underprices usage in the rest of the peak window within the billing period.
2. Unfair intraclass cost allocation, with those customers with demand diversity subsidizing those with more continuous usage.

Peak window demand charges can certainly elicit customer response and incentivize them

---

[47] The options that are available in practice depend on metering technology, which has evolved substantially over time. In the early part of the 20th century, TOU rates could be implemented with meters that operated on timers, where one track would record on-peak usage every day and another track would record off-peak usage every day. No distinction based on weekends or holidays was possible. By 1941, more sophisticated versions were available with remote controls that could switch the meters between tracks on command. Since that time, many more innovations have occurred to enable different types of time-varying rates. Three-period TOU rates became common for large industrial customers in France beginning in the 1950s. With advanced metering infrastructure and a sophisticated data collection and billing system, the possibilities are nearly endless. Even without AMI, simple TOU meters have long been available that track on-peak and off-peak usage based on programmed timers, which can exclude weekends and holidays from on-peak periods.

[48] Oklahoma Gas & Electric. (2018, June 18). *Standard pricing schedule: R-VPP variable peak pricing.* https://www.oge.com/wps/wcm/connect/c41a1720-bb78-4316-b829-a348a29fd1b5/3.50+-+R-VPP+Stamped+Approved.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-c41a1720-bb78-4316-b829-a348a29fd1b5-mhatJaA

to shift load from inside to outside that window.[49] Nevertheless, peak window demand charges share many of the faults of traditional monthly demand charges, just on a different scale. Once again, the key distinction is between the consistent and symmetric marginal incentive of a time-varying kWh rate and the arbitrary effects driven by the demand charge's ratchet.

A close examination of customer behavior reveals why energy-based prices are preferable to demand charges even within a peak window. In any system with significant customer diversity, a large number of customers will not have their individual peaks at the time of the system's peak. Still, it could be that a substantial number of customers peak at the time of the system peak. The proportion of one to the other matters if demand charges are to have a significant linkage to the system peak. Customers who are at risk of setting the individual peak for the demand charge face a high marginal price for consumption, but those who are not face a lower marginal price. This proportion will vary from service territory to service territory and over time as technology evolves.

Customer behavior under a peak window demand charge would likely even vary based on completely arbitrary factors. That could be whether certain customers are at the beginning of their billing period or whether a significant event that led a customer to incur a largely unavoidable peak (e.g., hosting a party during a peak window) happened before that very high-load time. This randomness can be entirely avoided. The fair and efficient solution is to continue to treat all consumption as marginal, a condition that is achieved by time-varying kWh rates.

In the absence of technology that automates response to changes in prices, the ratchet problem for peak window demand charges may be diminished by the inability of customers to respond accurately to its incentive structure. It is unlikely that people going about their daily lives can do more than respond to the broad incentives provided by either an on-peak kWh price in a simple TOU rate or a peak-window demand charge. In both cases, the easiest answer may very well be just "consume less during the peak window period."[50] This could mitigate the harm posed by the ratchet, but it also begs the question about the underlying rationale if there is no customer response.

A modest subset of residential customers may be able to respond to the next rule of thumb presented by a peak window demand charge: to operate as few end uses as possible simultaneously. Fully responding to the incentives posed by a demand charge requires

---

[49] See, for example, Stokke, A., Doorman, G., & Ericson, T. (2009). *An analysis of a demand charge electricity grid tariff in the residential sector.* (Discussion Paper No. 574). Statistics Norway, Research Department. https://ideas.repec.org/p/ssb/dispap/574.html

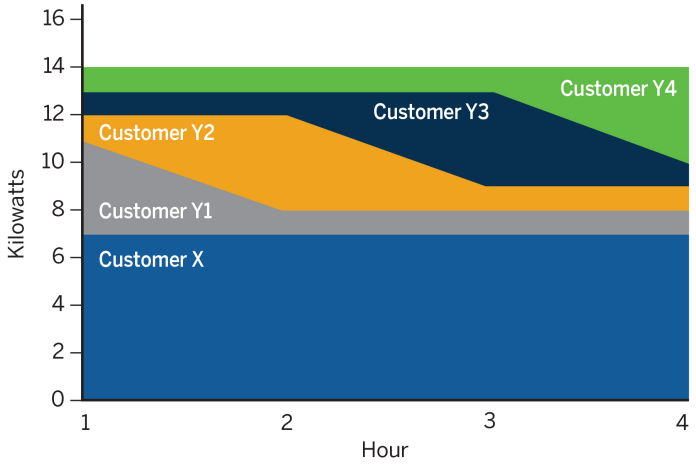
[50] For example, the Mid-Carolina Electric Cooperative has a three-hour peak window demand charge for residential customers. On the relevant page of its website, the peak window demand charge is labeled as an "on-peak charge." None of the advice given to manage this rate is specific to the actual working of a demand charge and could equally apply to a three-hour on-peak kWh rate. Mid-Carolina Electric Cooperative. (n.d.). *Rate structure.* http://www.mcecoop.com/content/rate-structure

customers to track their demand and know whether they are currently at risk of setting a high demand for the billing period — too much to ask of many residential and small business customers.

However, energy management technology, enabled by software and "supercharged" by on-site storage, will be able to adjust usage in a far more responsive manner than ordinary people could manage alone. Such energy management is likely feasible today for larger customers and could very well be widely feasible for smaller customers in the next few years. At least one company, Energy Sentry (http://energysentry.com/index.php), has developed a residential "demand controller" that automatically sheds less critical loads (water heaters, clothes dryers) when priority loads (microwaves, coffee makers, hair dryers) are activated. Such technology would allow customers to respond more effectively — *from their perspective* — to the incentives provided by a demand charge. But that is not to say that the overall efficiency of the electric system will be improved, since customer responses to demand charges do not typically optimize use of the system.

Peak window demand charges also create intraclass cost allocation problems, which are linked closely to the above efficiency concerns. Peak window demand charges still overcharge the low-load-factor customer and undercharge the high-load-factor customer. This is illustrated in the case of several smaller customers whose aggregate consumption adds up to the load of a single larger customer. Such a hypothetical is shown in Figure 5 for a four-hour peak period.

Figure 5. Customer load comparison illustrating ability to share capacity

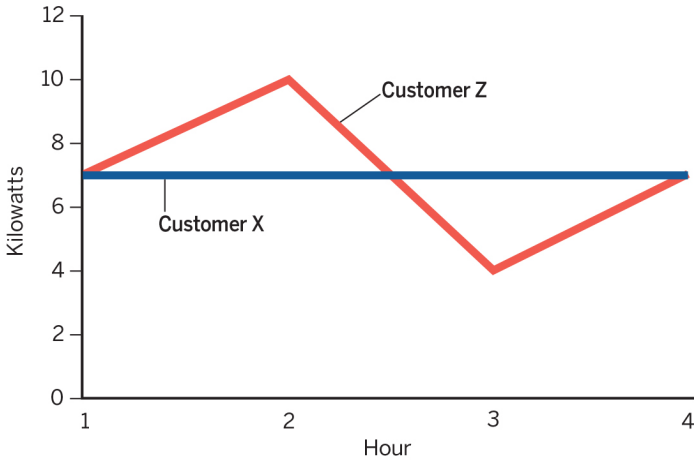


Customers Y1, Y2, Y3 and Y4 have, in the aggregate, the same load profile as Customer X. Each of the Y customers has a peak of 4 kW for a total billing determinant of 16 kW under a peak window demand charge. However, Customer X has a peak of 7 kW, which translates into a billing determinant of 7 kW under a peak window demand charge. This means that Customer X is charged less than half the amount that the Y customers are for the *exact same aggregate load pattern*. The four diverse customers can efficiently share capacity and should not be penalized by a price structure that fails to account for their diversity. Time-varying energy-based charges solve this problem.

Peak window demand charges, though an improvement on monthly NCP demand charges,

still come up short in the effort to send accurate information to consumers about peaks and other high-cost events. The occurrence of a peak cannot be known in advance, and, indeed, its timing depends in part on price structure. Shifting hours within a peak period does not necessarily lower the overall peak. Figure 6 is a comparison of two customers with equal kWh consumption in the peak period, one with a flat consumption throughout that period and another that varies.

Figure 6. Comparison of two customers



Compared with Customer X with a flat load pattern, Customer Z with the varying load pattern likely increases the chance of a system peak in hour 2, but by the same token likely decreases the chance of a system peak in hour 3. But the reverse can be said for Customer X compared with Customer Z: Customer X raises the likelihood of a system peak in hour 3 and decreases the likelihood of a system peak in hour 2. Advocates of demand charges consistently fail to explain why these types of discrepancies are justified by cost considerations.

Even well-designed TOU rates do not necessarily reflect critical peak times very well. For example, a four-hour weekday on-peak window for only the highest demand months will include around 200-400 hours annually. These will necessarily contain some days with higher peaks than others and only a limited number of hours that define utility capacity needs for reliability purposes at peak. Simple TOU rates do not distinguish in this regard between the moderate peaks (e.g., ordinary days in the summer) and the very highest peaks (e.g., extremely hot days in the summer). In short, the implication is that simple TOU rates do not provide a sharp enough incentive on actual peak days.[51] In any case, we are no longer bound to simple TOU pricing. Dynamic rates, including critical peak pricing, peak-time rebates, variable peak pricing and real-time pricing, all better address peaking issues because they provide higher marginal prices at the times of maximum system stress. By concentrating customer attention on the hours that actually drive costs, the more dynamic rates produce better results for the electric system and society.

By its very nature, a demand charge cannot present symmetric and consistent marginal incentives in the same way as a time-varying kWh charge. Compared to traditional demand charges, properly sized peak window demand charges have a better cost causation

[51] This is referred to as the needle-peaking problem in Crew & Kleindorfer, 1979, p. 186.

basis because they can be linked to the time periods that drive higher system costs. Daily as-used demand charges[52] applied to peak windows could be a further improvement on peak window demand charges, and, better yet, these peak window daily-as-used demand charges could fluctuate according to system conditions. However, this is only an improvement because it converges on the better solution, a system of time-varying kWh rates. Given the rate design possibilities that AMI offers, what reason is there to retain demand charges at all?

4. What Might Be Left for Demand Charges?

The foregoing demonstrates that the typical argument for demand charges, as used for generation, transmission and shared distribution capacity, is substantially flawed. Even so, we want to investigate if there are any circumstances, however limited, for which demand charges are an efficient rate design.

Some theorists have identified a different and, in our minds, much narrower set of rationales for demand charges. The case for time-varying rates relies substantially on the diversity of load and the lack of a direct relationship between individual customer peaks and the system peaks that drive costs. A diverse set of customers may, in the aggregate, create a predictable load profile much of the time. But what if this diversity goes away in an unpredictable manner? Or, for that matter, in a predictable one? Is there something about the causation of costs in special and *limited* circumstances that warrants charging for peak incurrences of short-term (e.g., 15-minute) demand for individual customers? To answer this question, we consider three cases that, on their faces, might present a marginal-cost justification for demand charges. The first is one that we have carved out from the beginning: capacity costs that are not shared, such as dedicated transformers and service drops, which we term "dedicated site infrastructure."[53] This illustrates some important issues relevant to any broader theoretical case for demand charges. The second is the cost associated with uncertainty in customer behavior. The third is timer peaks, a phenomenon where customers shift usage in response to hours with lower prices.

---

[52] RAP authors, writing with partners from Synapse Energy Economics, previously recommended daily-as-used demand charges for standby service to large combined heat and power customers, as an alternative to monthly standby demand charges. The purpose was to recognize that different combined heat and power customers would have scheduled and forced outages on different days and could share the same capacity to provide their standby service. This was certainly an improvement on monthly demand charges for such customers, but, in light of the progress made in metering and time-varying energy-based rate structures, there's every reason to think today that such time-varying energy rates are equally appropriate to customers with on-site generation. Johnston, L., Takahashi, K., Weston, F., and Murray, C. (2005, December 1). *Rate structures for customers with onsite generation: Practice and innovation*. National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy06osti/39142.pdf

[53] RAP has previously recommended a small transformer or site infrastructure demand charge for secondary voltage customers, particularly those customers with dedicated site infrastructure. See Lazar, J., & Gonzalez, W. (2015, July). *Smart rate design for a smart future*, pp. 53-54. Regulatory Assistance Project. https://www.raponline.org/knowledge-center/smart-rate-design-for-a-smart-future/

4.1 Dedicated Site Infrastructure

Dedicated transformers and service drops for individual customers are, by definition, not shared infrastructure. The relative importance of this category of cost will vary by customer class. Larger commercial and industrial customer classes, as long as they are taking secondary voltage service, will often have dedicated transformers for each customer or a dedicated transformer bank for customers taking three-phase power. Dedicated transformers will be rare for residential customers in urban and suburban areas, but single-family homes will almost always have a dedicated service drop. The largest industrial customers may have their own primary line (effectively serving as a dedicated service drop) or a dedicated substation (effectively serving as a dedicated transformer). In rural areas, each customer will typically have a dedicated transformer, at which point transformers are customer-specific site infrastructure.

For these customer-specific site infrastructure costs, there is no diversity of demand between the customer meter and point of connection with the shared system. As a result, individual customer NCPs are certainly relevant to the sizing of these components. One might conclude from this that a demand charge can provide a reasonable pricing incentive here. The time period for such a demand charge should have nothing to do with a shared peak since there is no sharing of the infrastructure. Nor should it be limited to peak windows since the peak for an individual customer could occur at any time. The cost of these components may be no more than about $1/kW/month, a fraction of typical demand charges.[54]

There are also other ways of efficiently pricing this category of costs. A similar set of customer incentives may be presented by a connected load charge for a set amount of local capacity. Such a connected load charge can help with efficient sizing, but only if it's accompanied by a fee for overages or the automatic tripping of circuits when demand would cause an overage. Even then, a connected load charge provides no incentive for customers to manage their usage efficiently; that is, there are no cost savings to be gained by keeping their demand below the level of the predetermined connected load. A charge that establishes the relationship of the customer's individual peak demand to the sizing of these components might, however, give the customer some incentive to minimize peaks.

It is worth examining this issue at the level of engineering and planning. What type of customer behavior would minimize the risk of transformer overload and degradation? Or what type of customer behavior would allow utilities to size dedicated transformers more efficiently?

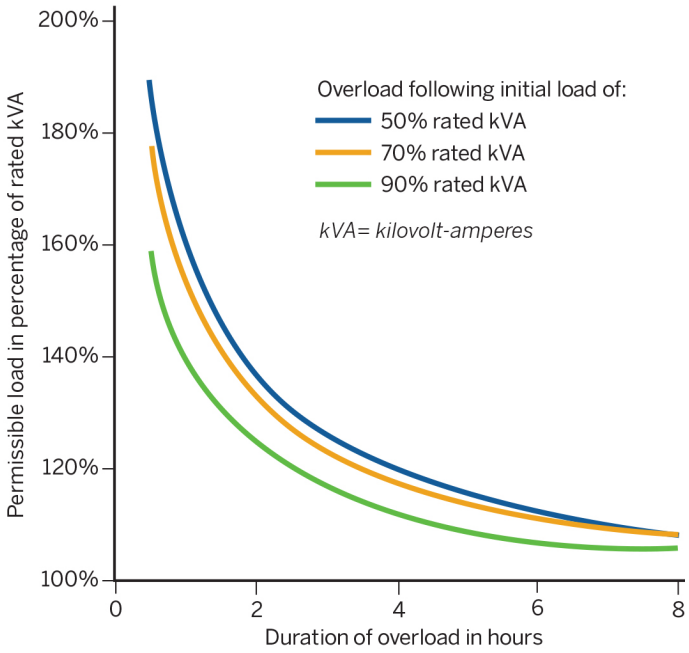
Capacity ratings for the different elements of the electric system are set with many

[54] Seattle City Light, for example, has a large general service rate with specific charges for transformer investment; these are $0.27/kW/month. Seattle City Light. (n.d.). *City Light rates.* https://www.seattle.gov/light/rates/summary.asp

engineering limits in mind. Many of the most important considerations revolve around the heating — and overheating — of components, particularly transformers and conductors. This has a number of different implications. For example, effective delivery capacity can be higher in the winter than the summer or higher in the cool nighttime than during the sunny daytime. The capacity ratings for individual system elements are for sustained loads in typical conditions, but loadings can exceed those ratings on a regular basis without necessarily incurring significant damage. As Tom Short colorfully puts it, a conductor rated 480 amps "will not burst into flames at 481 [amps]."[55]

Figure 7[56] demonstrates the maximum overload that a transformer can take without shortening its operating life, by examining two primary variables: (1) the initial load prior to any overload and (2) the duration of an overload. If a transformer has had light loads (50% of its rating), it can sustain a short-term overload of nearly 190% or a four-hour overload of just over 120%.

Figure 7. Permissible transformer overloads for varying periods



Source: Bureau of Reclamation. (1991). *Permissible Loading of Oil-Immersed Transformers and Regulators*

The important question then is what kind of rate design incentivizes optimal customer behavior with respect to this equipment. Panagiotis Andrianesis and Michael C. Caramanis have developed an algorithm for dynamic nodal locational marginal costs for distribution systems that offers an intriguing approach to pricing for these customer-specific facilities. For line transformers, the pricing formula is a real-time price per unit of energy that follows the transformer thermal response dynamics, which is essentially the temperature of the cooling oil in each transformer.[57] Similarly, a critical peak energy charge could apply for the few hours per year when a transformer is

[55] Short, T. A. (2004). *Electric power distribution handbook*, Section 3.5, p. 140. CRC Press.

[56] Bureau of Reclamation. (1991). *Permissible loading of oil-immersed transformers and regulators.* https://www.usbr.gov/power/data/fist/fist1_5/vol1-5.pdf

[57] Andrianesis, P., & Caramanis, M. (2019). *Distribution network marginal costs: Part 1, A novel AC OPF including transformer degradation.* arXiv. https://arxiv.org/abs/1906.01570

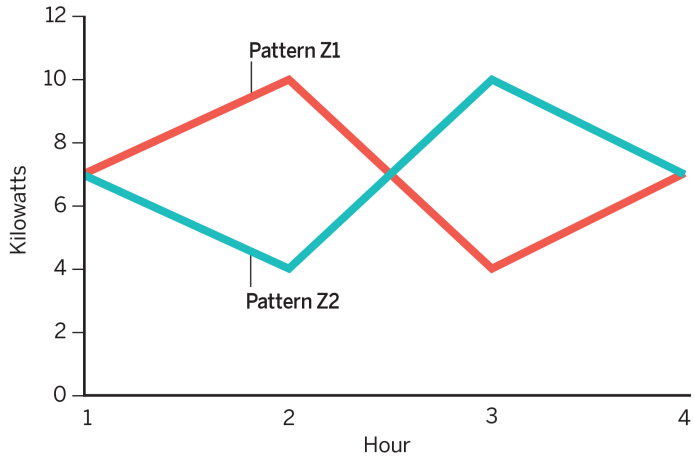
stressed but would require real-time monitoring and pricing to be applied on a transformer-by-transformer basis.

This type of short-run marginal cost pricing does not resemble a demand charge and has the virtue of linking closely in time the incurrence of high marginal costs to the prices charged. These approaches are quite sophisticated and could be costly to administer. To achieve a rate that is more feasible now, a simpler structure would be necessary. A daily-as-used demand charge or a traditional monthly demand charge, based on 15-minute or 30-minute peaks, could certainly discourage the extremely high short-term peaks that would damage a transformer. Those options might not do enough, however, to discourage a sustained, multihour overload.

4.2 Risks of Customer Variance at Peak Times

Load diversity isn't static and can fluctuate in ways that are both predictable and unpredictable. Predictable changes often occur around the weather, one of the few variables that simultaneously affects all customers in a given area. Regarding unpredictable changes, consider a simple hypothetical illustrated in Figure 8.

Figure 8. Two hypothetical load patterns randomly chosen by customers



If there are 10 "random-load" customers who flip a fair coin to determine whether their load profile corresponds to either Z1 (heads) or Z2 (tails) in Figure 8, the average load in each hour across a large number of trials will be 70 kWh.[58] However, system planning must not only deal with the expected average load but rather the chances of higher load. Unfortunately, in any given trial of this scenario, the probability of five heads and five tails — leading to a demand of 70 kW in every hour — is only 24.6%. There is a small but nonzero chance that every customer gets either heads or

[58] In this illustrative example, we consider each customer to have a flat load within each hour. This means that kW and kWh are largely interchangeable as units. Similar examples could, however, be constructed with demand varying in smaller increments (e.g., 30, 15 or 5 minutes), and similar results could be obtained.

tails, leading to a 0.2% probability of a peak load of 100 kW. The full spectrum of potential results for this hypothetical scenario with 10 random-load customers is shown in Table 1.

Table 1. Peak load and probabilities for 10 random-load customers

Coin flip result	High load: Number of customers	Low load: Number of customers	Peak load (kWs)	Probability
10 heads or 10 tails	10	0	100	0.2%
9 heads or 9 tails	9	1	94	2.0%
8 heads or 8 tails	8	2	88	8.8%
7 heads or 7 tails	7	3	82	23.4%
6 heads or 6 tails	6	4	76	41.0%
5 heads and 5 tails	5	5	70	24.6%

The cumulative odds of a peak of 88 kWh or higher is 10.9%, and a peak of 82 kWh or higher is 34.4%. In this hypothetical scenario, it is clearly beneficial to have customers flatten their load curves to 7 kW every hour within this time period. Table 2 shows the range of possible results and associated probabilities for six random-load customers corresponding to either pattern Z1 or Z2 and four flat-load customers with a demand of 7 kW in each hour.

Table 2. Peak load and probabilities for six random-load and four flat-load customers

Coin flip result	High load: Number of customers	Low load: Number of customers	Flat load: Number of customers	Peak load (kWs)	Probability
6 heads or 6 tails	6	0	4	88	3.1%
5 heads or 5 tails	5	1	4	82	18.8%
4 heads or 4 tails	4	2	4	76	46.9%
3 heads and 3 tails	3	3	4	70	31.3%

The risk of a peak of 88 kW or higher drops from 10.9% to 3.1%, and the risk of a peak of 82 kW or higher drops from 34.4% to 21.9%. If the customer choices are uncorrelated, this type of risk goes down as the number of customers increases.[59] However, if the customer choices are correlated, between hour 2 and hour 3 in this hypothetical, the risk does not necessarily decrease with a higher number of customers.

[59] The ratio of the variance to the expected total decreases in proportion to the square root of the number of customers.

This simple example is the essence of the argument made by Michael Veall.[60] He demonstrates that, for a given level of average customer demand during a peak, higher variance customers lead to a risk of higher peaks, particularly if they are correlated. This, in turn, results in a need for higher capacity planning margins. Veall constructs a detailed economic model of optimal peak period pricing. He states that the traditional monthly demand charge does not reasonably address this issue, but rather a peak window demand charge can serve as marginal price on a customer's variance. He notes additional caveats: "If there are many small users with uncorrelated demands, the effects of an individual user's variation on total system variation will be small. But if users are large or their demands are correlated, variance charges are important."[61] Finally, Veall's result demonstrates that, if a peak window demand charge is to be imposed, it should be paired with an on-peak kWh rate.[62] The logic around risk and Veall's theoretical model present an argument for a peak window demand charge that is substantially different from those that utilities put forward. And again, we see a more defensible justification for peak window demand charges for larger-volume customers. But the key question of correlations and *levels of risk* has been neglected in the discussion around demand charges and is only a theoretical possibility in Veall's model. Furthermore, Veall's model does not consider the possibility of more granularly dynamic time-varying kWh rates.

Marcel Boiteux, the influential French economist and executive for Électricité de France (EdF), does discuss risk and uncertainty in a 1952 paper, written jointly with his colleague Paul Stasi.[63] When it comes time for tariff design, Boiteux and Stasi describe two different zones of the shared electric system: (1) the "collective network" and (2) the "semi-individual network, whose capacity depends particularly on the uncertainties of consumption of each customer."[64] With respect to the collective network, they find that the "uncertainties of individual consumption" are small enough to be ignored.[65] And, finally, their analysis of the "semi-individual network" is dominated by risk and the irregularities of individual customer's loads. This leads them to a justification for a complex system of subscription-based contract demand charges, with higher prices for contracted demand in

[60] Veall, M. (1983). Industrial electricity demand and the Hopkinson rate: An application of the extreme value distribution. *The Bell Journal of Economics, 14*(2), 427-440.

[61] Veall, 1983, p. 429.

[62] Veall, 1983, p. 431. Veall notes that this on-peak kWh price could, in principle, even be negative, which would be a curious result.

[63] Boiteux, M., & Stasi, P. (1964). The determination of cost of expansion of an interconnected system of production and distribution of electricity. In J. Nelson (Ed. & Trans.). *Marginal cost pricing in practice.* Prentice Hall. (Original work published in 1952).

[64] Boiteux & Stasi, 1964, p. 117.

[65] Boiteux & Stasi, 1964, p. 117

peak periods and lower prices in other periods. However, Boiteux and Stasi offer little but generalities as to the demarcation of the semi-individual network:

> The extent of the zone within which the uncertainties of individual demands have a very marked influence on the collective cost is greater in proportion to the irregularity of the demands considered, and to the correlations among these demands. This extent depends also on the density of consumption, for the number of customers supplied from a given node plays an important role in the "reduction of uncertainties."[66]

Based on these considerations, Boiteux and Stasi largely describe the generation and transmission system (150-220 kV) as the "collective network" and the distribution system (15-60 kV) as the "semi-individual" network.[67] They are discussing these issues in the context of the then-new Tarif Vert for high voltage industrial customers, and the discussion could be read in a manner that is limited to those customers. This could mean that the dividing line for the semi-individual network could vary by the size of customer. Whether residential customer fluctuations are correlated in a significant and pertinent way is another empirical question Boiteux and Stasi do not address. It is unclear whether the irregularities of individual residential customers would ever be significant enough to matter at a level higher than a shared transformer.

4.3 Timer Peaks

Michael A. Crew and Paul R. Kleindorfer (1979) raise another area where a demand charge could theoretically be efficient, which they describe as the secondary preferences of customers, given the structure of time-varying rates, and the shifting of demand to notionally off-peak times. This is colloquially known as a timer peak. This occurs if customers increase their usage substantially during hours with low rates or more specifically right at the time when low-priced hours begin.[68] In the worst-case scenario, TOU rates can theoretically just shift the system peak without reducing it if enough usage is shifted to hours with low prices. The same is true of coincident peak window demand charges. This outcome can be avoided by managing the number of periods in the rate, the hours covered by each period and the relative prices. One utility has designed a TOU rate in which each customer chooses a three-hour peak period of 4-7 p.m., 5-8 p.m. or 6-9 p.m. All of these customers have 6-7 p.m. in their peak period; two-thirds of them have 5-6 p.m. and 7-8 p.m. in their peak period; and one-third have either 4-5 p.m. or 8-9 p.m. in their

[66] Boiteux & Stasi, 1964, p. 123

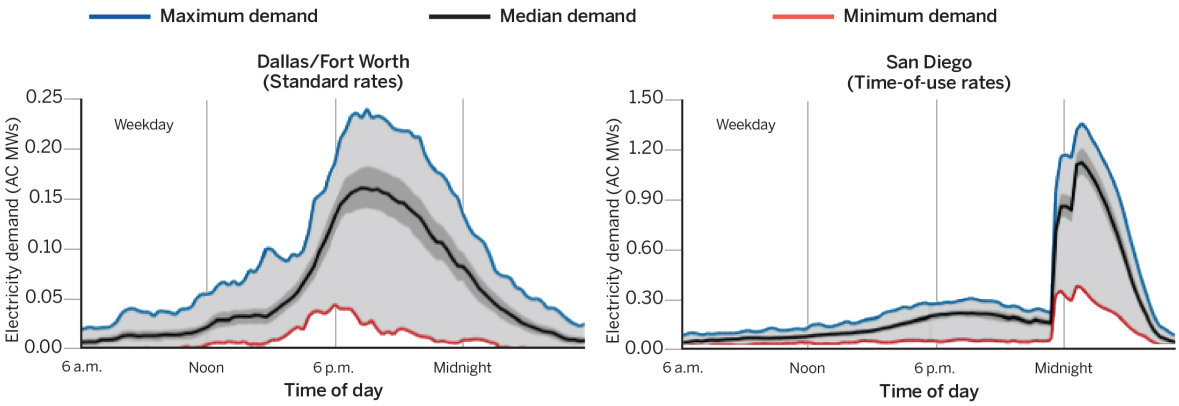
[67] Boiteux & Stasi, 1964, p. 110.

[68] Bonbright mentions this as a possible objection to time-of-use rates. See Bonbright, 1961, p. 362, fn 23: "In Chapter 10 of his book already cited in footnote 10, Davidson suggests this type of rate [time of day and time of season] as preferable to the familiar Hopkinson-type rate. But among the objections to it is the danger that its sharp breaks will create surges in the loads imposed on a power station or on a distribution line."

peak period. This provides the utilities with the most powerful pricing signal during the most likely peak hour but substantial peak reduction in the adjacent hours.[69] Another simple solution is to apply different hours to different classes or subclasses. For example, single-family residences may have their tariff shifted one hour earlier than apartments, or secondary general service one hour later than primary general service.

If the more general system peaks are not impacted significantly enough by this phenomenon to warrant changing the structure of the rate, more granular and local issues can theoretically arise. Figure 9 shows a set of results from a San Diego Gas and Electric rate for electric vehicles, where the "super off-peak" rate begins at midnight.[70]

Figure 9. Real-world illustration of timer peak with EV charging on TOU rate



Source: Jones, B., Vermeer, G., Voellmann, K., & Allen, P. (2017). *Accelerating the Electric Vehicle Market*

This is a rational customer response to a TOU rate, at least for specific end uses. If an electric vehicle is parked at home in the evening and will not be used again until morning, then the customer has a significant amount of flexibility to choose when charging will begin. The very beginning of the lowest price period is an obvious time to start charging. While this may not present an issue at the generation and transmission level, assuming midnight remains an off-peak period, bunching of EV charging could lead to issues at the more local level.

Again, thinking about a line transformer helps focus the analysis. If five single-family homes are served by one shared transformer and all five of those homes have EVs that start charging at midnight, then impacts at the line transformer level are a possibility. Furthermore, what if those houses have other timed usage that starts at the beginning of the lowest price period? Sending a secondary price signal that discourages households

[69] Salt River Project. (n.d.). *SRP EZ-3 price plan*. https://www.srpnet.com/prices/home/ez3.aspx

[70] Jones, B., Vermeer, G., Voellmann, K., & Allen, P. (2017). *Accelerating the electric vehicle market*, p. 16. M.J. Bradley & Associates. https://www.mjbradley.com/sites/default/files/MJBA_Accelerating_the_Electric_Vehicle_Market_FINAL.pdf

from turning on all of their major end uses at midnight could be effective. A daily-as-used demand charge can send such a signal if it were applied across 24-hour periods. A more traditional monthly demand charge, however, will send that signal in a much more attenuated way. A connected load charge based on a contract demand for the cost of connection, with fees for overages, similar to the Électricité de France Tempo rate for residential customers,[71] would send a similar price signal as well.

There are other ways to deal with this phenomenon besides price signals with demand-charge features. The beginning of the off-peak period with the lowest price could be staggered for different customers, for example, beginning at 10 p.m. for one-third of customers, midnight for another third and 2 a.m. for the last third. For customers with long-duration controlled loads, like water heaters or electric vehicles, this would be easy for the customer to manage and beneficial to the utility. To maximize the benefits of such an approach, one would need to have a relatively even split among those three options for the customers on each shared transformer. Load management programs and smart devices could deal with this type of issue as well. For some loads, particularly water heating and EV charging, we anticipate advanced devices that will enable the utility to manage loads to minimize costs and enable customers to benefit from even lower off-peak rates for enabled devices.

One could even question how much of a problem this poses to the longevity of the shared transformers in question. The ambient temperature has almost always cooled off by midnight, and several hours of low or moderate loads could allow the transformer to cool from significant levels of usage during the day or early evening. In certain circumstances, it could be more convenient and cost-effective to upgrade any potentially affected transformers, particularly where multiple water heaters or EV chargers are served from a single transformer.

5. Conclusion

Demand charges, of either the traditional monthly NCP or peak window variety, are not efficient, as a general matter, for shared system capacity costs because:

- For the vast majority of customers, any peak reduction signal in a traditional monthly demand charge is weak and inaccurate.
- Traditional calculations for demand charges have included far too many costs as demand-related. Ideally, utility commissions will adopt a new time-based classification

[71] Electricite de France. (n.d.). *Tarif Bleu: Regulated sale tariff for electricity*. https://particulier.edf.fr/en/home/energy-and-services/electricity/tarif-bleu.html

and allocation framework for generation, transmission and shared distribution costs.[72] Failing that, the numerous energy benefits from capacity investments should be properly accounted for — that is, reflected in energy, not demand, charges.

- Simple TOU rates are superior to peak window demand charges in their own right, but AMI enables time-varying energy charges, such as critical peak pricing, peak-time rebates and variable peak pricing, that much more accurately target times of system stress and reward end users for shifting their loads to off-peak times.

Although we have shown the significant downsides to using current forms of demand charges, in very limited circumstances there might be cost- and efficiency-related justifications for certain types of demand charges. But such charges would be significantly lower than those prevailing for industrial customers in the United States today. Dedicated site infrastructure is a small portion of utility system costs, and typical demand charges would not necessarily provide an optimal signal to control these costs. The primary concerns around timer peaks are almost certainly limited to local infrastructure.

As for the general risk of customer variance and correlation, little work has been done to investigate the statistical bases of this more sophisticated case for demand charges. We think that it is unlikely that such an analysis would find that a substantial demand charge would be fairer or more efficient than time-varying energy charges. Lastly, there is a better case for demand charge-like structures for large customers, who are more likely to have significant dedicated site infrastructure. One might also argue that high variance at peak times among these customers has a more significant chance of influencing the overall system peak. Any such demand charges may not look like Hopkinson rates and would likely be only a second-best solution to a sophisticated system of time-varying energy charges.

The economic and regulatory principles that underlie these judgments are not new. The inescapable essentials of microeconomic theory are at work here. Boiteux, Bonbright and Kahn follow these principles and theories, as do the other scholars and practitioners we cite. In 1964, Paul Garfield and Wallace Lovejoy[73] also stepped into the fray. They converted principles of economic efficiency and fairness into straightforward criteria for assessing the merits of cost allocation methods and rate designs for generation and delivery capacity costs:

- All utility customers should contribute to capacity costs.
- The longer the period of time that customers preempt others' use of capacity, the more they should pay for the use of that capacity.

[72] Lazar et al., 2020.

[73] Criteria adapted from Garfield, P., & Lovejoy, W. (1964). *Public utility economics,* pp. 163-165. Prentice Hall.

- Any service that makes exclusive use of a portion of capacity should be assigned all of the costs for that portion of capacity.
- The allocation of capacity costs should change gradually with changes in the pattern of usage.
- More capacity costs should be allocated to on-peak usage than off-peak.
- Interruptible service (or other forms of utility restrictions and control) should be allocated less in capacity costs as the degree of restriction increases.

Only time-varying energy charges can meet all of these objectives simultaneously. Demand charges for shared costs are demonstrably less efficient and less equitable than they.

Energy Solutions for a Changing World

Regulatory Assistance Project (RAP)®
Belgium · China · Germany · India · United States

50 State Street, Suite 3
Montpelier, Vermont 05602
USA

802-223-8199
info@raponline.org
raponline.org