memo

To: NEPOOL Markets Committee

From: Market Development and Business Architecture & Technology

Date: December 7, 2015

Subject: FCM Zonal Demand Curve Methodology – *Revised Edition*

This memorandum discusses the ISO's proposed method for developing capacity demand curves for the annual Forward Capacity Auctions (FCA). Building on a set of broad design principles, it describes a practical method for specifying capacity demand curves that can be readily applied to both the current and potentially different future capacity zone configurations. The proposed method is built on a clear engineering-economic foundation, and is designed to procure capacity cost-effectively relative other potential demand curves.

This revised edition incorporates and supersedes the contents of the memorandum on this topic provided to the Markets Committee on November 3rd, 2015. This revision adds a detailed discussion of demand curves for export-constrained capacity zones, and provides an expanded explanation of cost-effective capacity procurement between zones.

Importantly, this analysis indicates that modifications to the existing system demand curve are desirable in conjunction with the zonal demand curves. We explain these modifications to the system curve in detail, and discuss some of the problems that arise without these modifications. In essence, the system-level and zonal demand curves should be developed in an integrated manner, since they jointly determine how much capacity clears in each portion of the system.

The proposed method will also enable the FCA to incorporate certain improvements to the clearing process between zones. Under prior zonal design proposals, if an import-constrained zone cleared at a higher price than the system, the aggregate capacity cleared would fall short of the system's sloped demand curve. In contrast, the new method will enable the rest-of-system zone to clear on the system-level demand curve if an import or an export zone clears on a zonal demand curve at a different price than the rest-of-system. This memorandum also explains these clearing process improvements, and how they are enabled by the proposed demand curve design.

The ISO anticipates discussing this subject with stakeholders over the next several months. We welcome additional questions and feedback on these issues.

Design Principles

In some respects, demand curves for capacity serve a purpose similar to demand curves for other goods and services. They characterize how much more consumers will buy when the price is low, and how much less they will buy when the price is high. The practical question before us is: At different prices, how much more or less should be purchased in the capacity market?

The ISO's proposed answer to this question starts with several design principles. These principles are:

1. *Reliability.* The demand curves should be expected to procure sufficient capacity, both in the system and in any modeled capacity zone, to enable the ISO to meet its reliability planning obligations. In practice, this is assessed on the basis of meeting the '1-in-10' loss of load expectation (LOLE) reliability standard, on average (over time).

2. *Sustainability.* The capacity market's clearing prices should remunerate investment in capacity at a level sufficient to attract new entry when needed. This means that the capacity market's equilibrium needs to exhibit prices that average at least the (true) net cost of new entry; otherwise, the FCA would not clear sufficient supply to meet the reliability objective (Principle 1).

3. *Cost-Effectiveness.* Whenever clearing prices differ among capacity zones, the zonal demand curves should be designed to minimize the bid-cost of procuring capacity overall. That is, capacity purchases should be allocated across zones cost-effectively, given each auction's prices, while meeting the overall reliability and sustainability objectives (Principles 1 and 2).

In addition to these three design principles, there are other benefits of using sloped demand curves in the FCA. One is that, relative to procuring fixed quantities, sloped demand curves tend to reduce volatility in auction clearing prices over time. That implies less year-over-year variation in revenues (for capacity suppliers) and expenditures (for capacity buyers). A second benefit is that sloped demand curves can significantly attenuate the incentive for a participant to exercise market power in the FCA. This is because sloped demand curves reduce the effect on the market clearing price if a seller raises (or, in the buyer-side case, a buyer lowers) a resource's bid price in the FCA.

An important further consideration is design 'robustness' to alternative zone configurations. One of the ISO's practical objectives is capacity demand curve rules that will work – in the sense of satisfying the design principles and benefits noted above – even if the FCA is conducted using a different zonal configuration at some point in the future. A cavalier set of curves may seem appealing for one zonal configuration, but if the underlying methodology is not designed carefully so as to handle alternatives, it may leave little confidence that it will work in future years.

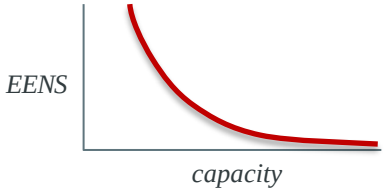
In developing an improved method for determining demand curves to be used in each FCA, the ISO has not elevated some of these principles or benefits above the others. Rather, we emphasize the three enumerated design principles in the analysis that follows because they have precise interpretations that are a valuable guide for developing demand curves for the FCA. The balance of this memo explains how.

Economic Foundations

The FCA can be viewed as making trade-off between two basic considerations. At a high-level, the benefit of procuring more capacity is that it reduces energy demand that may go unserved (sometimes called 'lost load'). However, procuring more capacity is costly. Capacity demand curves can be viewed as a means to make an economic choice about how this tradeoff between the costs and benefits of procuring more or less capacity should be resolved.

■ **Capacity Avoids Lost Load.** To evaluate the potential lost load that should be expected with a given level of installed capacity, the ISO's reliability planning models calculate a performance metric known as the 'expected energy not served' (EENS).[1] EENS is measured in MWh per year, and depends on (among other things) the amount of capacity installed on the system and in each constrained zone.

EENS is an informative measure of how frequently an incremental unit of capacity should be needed to avoid lost load. Graphically, the relationship between EENS and capacity looks like this:



Intuitively, when the system has low levels of capacity, procuring an incremental MW of capacity tends to have a big reliability benefit because it reduces lost load in many hours per year. At the opposite extreme, when the system has high levels of capacity, an incremental MW of capacity generally has little reliability benefit: it is rarely needed and reduces lost load in very few (if any) hours per year.

The ISO's approach to developing capacity demand curves is based, in part, on how expected lost load – more specifically, EENS – is impacted by procuring different capacity levels (zonally and system-wide). To explain this approach, we first explain how to interpret the FCA as a cost-minimization problem using EENS, rather than using a 'fixed' capacity requirement.

■ **The FCA as a Cost-Minimization Problem.** At a conceptual level, the capacity auction's design can be thought of as a cost-minimization problem. Until FCA 8, it sought to minimize the costs of procuring capacity, subject to a fixed quantity requirement. Without a fixed quantity requirement (FCA 9 and subsequently), the FCA instead makes a tradeoff – albeit implicitly – between the cost of procuring capacity and an implied 'benefit' of each additional MW procured. To justify a particular set of demand curves, it is insightful to view the FCA as a minimization problem in this new context, and to be more explicit about what is the incremental 'benefit' of each MW procured.

[1] Also known as 'expected unserved energy' (EUE). We use the two term synonymously.

In an FCA without a fixed capacity requirement, one can conceptualize the FCA's clearing objective as a cost minimization problem that minimizes both the total bid-based costs of capacity procured and the total 'cost' of the expected energy not served. To explain this more precisely, a formula will help. Let i represent a resource, and b_i denote its bid price (or offer price) in the FCA. Further, let q_i represent the quantity (i.e., the CSO MW) that resource i is awarded the FCA. (Thus, if a resource is not awarded a CSO at all, its value of q_i is zero.) Minimizing the total bid-cost of capacity and the total 'cost' of EENS can be expressed succinctly as:

$$\text{minimize}_{\{q_i\}} \quad \underbrace{\sum_i b_i \cdot q_i}_{\text{Cost of capacity (as-bid)}} + \underbrace{PF \cdot EENS(Q_{SYS})}_{\text{Cost of lost load}} \qquad (1)$$

In this expression, the first term (the summation) is the total bid-cost for all CSO MW's awarded. The second term can be interpreted as the total cost associated with the expected energy not served when total cleared capacity is Q_{SYS}. Note that—for the moment—we've assumed in this formulation an unconstrained system with no capacity zones. (We extend this to capacity zones further below).

Conceptually, this minimization problem can be viewed as a 'penalty factor' model, in which an incremental 'cost' (the penalty factor) of *PF* is assigned to each MWh of expected energy not served. Unlike having a fixed capacity requirement, an auction built on this foundation will minimize the total bid-costs of capacity (the first term) and expressly account for the additional reduction in unserved energy from procuring incremental capacity. In general, market designs based on penalty-factor models such as (1) are a familiar, practical means to send market price signals and procure supply cost-effectively when there is both a cost to acquire a good (in this case, capacity) and a cost associated with some form of shortage or unmet demand.[2]

To be clear, the expression in (1) serves a conceptual purpose – is it not how the FCA's market clearing engine currently works, nor a complete mathematical representation of how the ISO proposes to clear the FCA in the future. Nonetheless, this conceptually simple cost-minimization idea conveys some important economic points for how capacity demand curves should be developed for the FCA.

■ **From Penalty Factors to Demand Curves**. In order for the FCA to be cost-effective, it must procure capacity so that the marginal cost of another unit of capacity is equal to the marginal reliability benefit that capacity provides. That is, the FCA should clear a level of capacity such that

$$\textit{Marginal Capacity Cost} = \textit{Marginal Reliability Benefit} \qquad (2)$$

Although the reliability benefits of capacity can have many dimensions, we will focus on its benefit in the form of reducing expected energy not served (the resource adequacy problem). In that context, applying this cost-benefit principle using insights from the penalty-factor model in (1) provides a

[2] For example, the ISO's market model for real-time energy and reserves achieves its cost-minimization objective using a closely-related penalty factor model. The principle difference from (1) is that the real-time reserve pricing model uses a fixed reserve requirement, instead of a smoothly-decreasing $EENS(Q_{SYS})$ function. As seen shortly, using $EENS(Q_{SYS})$ yields a demand curve that declines smoothly with additional supply.

reasonable economic framework for developing capacity demand curves. To see how, we'll next examine both sides of the expression in (2), and then connect these concepts to the supply and demand curves used to clear the FCA.

On the left-hand side of (2), the marginal capacity cost is determined by the bid-price submitted by the marginal (that is, price-setting) capacity supply resource in the FCA. Nothing new there, and nothing about the development of capacity demand curves should change the interpretation of capacity supply curves. In a capacity auction for an unconstrained system (*i.e.*, without any zones), the marginal capacity cost at each level of capacity is simply the aggregate supply curve in the FCA.

On the right-hand side of (2), the reliability benefit of procuring capacity arises from reducing the total cost of lost load. That has two components. The first is the impact on expected energy not served of another unit of capacity. We will refer to this as the *marginal reliability impact* (MRI) of capacity. Stated mathematically, for an unconstrained system without capacity zones:

$$\textit{Marginal Reliability Impact} = \frac{d}{dQ_{SYS}} EENS(Q_{SYS}) \qquad (3)$$

The second component is the incremental 'cost' (the penalty factor) assigned to each MWh of expected energy not served. Stated more precisely, the cost-minimization model for the FCA in expression (1) corresponds to the following marginal reliability benefit:

$$\textit{Marginal Reliability Benefit} = -PF \times \textit{Marginal Reliability Impact} \qquad (4)$$

In words, the marginal reliability benefit is determined by the reliability impact of procuring another increment of capacity that reduces expected energy not served (in MWh per year), evaluated at (*i.e.*, multiplied by) an incremental 'cost' of *PF* assigned to lost load (in $ per MWh).[3]

You can see where this is headed. Conceptually, if demand curves for capacity are specified on the basis of its marginal reliability benefit, using the formula in expression (4), then running the FCA will procure a level of capacity that satisfies the familiar benefit-cost logic in expression (2). This occurs because the capacity auction's market mechanism clears at the point where the capacity supply curve intersects the capacity demand curve. That is, the FCA will procure capacity where its marginal capacity cost equals its marginal reliability benefit.[4]

Practicalities

The ISO's proposed approach to developing sloped capacity demand curves is based on the simple logic that the cost of capacity should be reasonably commensurate with its marginal reliability benefit. As will be shown presently, this economic foundation can be readily used to build zonal demand curves.

[3] The negative sign in expression (4) keeps the marginal reliability benefit positive. This is because the marginal reliability impact in (3) is negative: Additional capacity reduces expected energy not served.

[4] We simplify, though the point is general. In practice, capacity supply bids/offers can be 'lumpy' (non-rationable) and, in certain situations, this may result in these marginal conditions holding approximately rather than exactly. That does not undermine the logic for specifying demand curves as described here, however.

As a practical matter, applying the logic in expressions (3) and (4) to obtain capacity demand curves requires a reasonable basis for determining the penalty factor, and a practical process for evaluating the marginal reliability impact of procuring additional capacity. We touch on each of these two issues next, and provide additional detail in subsequent sections.

■ **Deriving a Penalty Factor.** How should a penalty factor on lost load be interpreted, and how should it be determined? Broadly answered, a high value of *PF* means that the region is willing to pay a great deal to avoid unserved energy demand, and the FCA should tend to procure a great deal of capacity to minimize this possibility. A low value of *PF* would result in the FCA procuring less capacity, though how much less will depend on the offer prices of capacity suppliers.

As a general matter, the ISO does not propose to assume a specific value for consumers' incremental 'cost' of expected energy not served; this is empirically difficult to ascertain with any confidence. Rather, the ISO proposes to *derive* (not assume) a value for the penalty factor *PF* that is (just) high enough to simultaneously satisfy the Reliability and the Sustainability design principles described at the outset of this paper. In that way, the penalty price will be set at the level just necessary to produce outcomes consistent with the reliability planning standards, on average and over time, and to induce new entry when necessary. We present a simple method to do this in detail further below.

It is worth noting that *any* capacity demand curve necessarily has – at least implicitly – *some* value assigned to the incremental 'cost' of lost load. In past design efforts, this value was implicitly determined by the curve's location and slope.[5] In contrast, the ISO's revised approach explicitly acknowledges this parameter, and the role it plays in affecting capacity demand curves. That is, instead of leaving the incremental 'cost' of lost load as an unstated parameter, this approach adopts a (derived) value calculated to procure capacity levels consistent with the region's reliability planning requirements, on average and over time.

■ **Evaluating the Marginal Reliability Impact.** Although the marginal reliability impact formula in expression (3) may seem complex or novel, it is neither. It has a simple interpretation as the change in expected energy not served with respect to capacity procured. This is a smoothly increasing curve. Indeed, the ISO's existing reliability planning models are able to produce the EENS curve and its gradient (the MRI) presently, both for an unconstrained system and with the capacity zones modeled in the FCA. These curves are obtained using the system parameters and inputs that are currently used

[5] An example may help. Using round numbers: Suppose that at NICR a (hypothetical) system demand curve specifies a price of $120,000 / MW-year (*i.e.*, $10 per kw-month). Suppose further that the marginal reliability impact of another increment of capacity at NICR is a 6 MWh per year of EENS reduction. The implied penalty factor this demand curve assigns to lost load at NICR is $120,000 / 6 MWh = $20,000 per MWh annually. Stated in other words, since this (hypothetical) demand curve does not procure another increment of capacity (beyond NICR) when additional capacity is offered at a price of $120,000 per MW-year and the next increment would reduce EENS by 6 MWh, the demand curve implicitly assigns an incremental 'cost' of avoiding lost load, at the margin, of $20,000 per MWh annually. Note that, as a technical matter, the implicit penalty factor on EENS may not be constant for all capacity levels under some demand curves (it is constant under the ISO's method).

to calculate various ICR-related values for each auction, and that are vetted annually with stakeholders.[6]

■ **Implications.** There are three main implications of this analysis for demand curves:

1. There is an engineering-economic foundation for developing FCA demand curves based on the marginal reliability impact of capacity. Doing so will enable the FCA, by clearing the auction where capacity supply meets capacity demand, to procure capacity levels that seek to balance the cost of capacity and the cost of avoiding lost load.

2. A reasonable basis for assessing the marginal reliability benefit of capacity is the product of two terms: (a) The impact of incremental capacity purchases on EENS, which the ISO's existing reliability planning models can calculate, and (b) a penalty rate for each MWh of EENS. As explained further below, an appropriate value for the penalty rate can be derived using the Reliability and Sustainability central design principles.

3. This economic foundation is not specific to capacity zones, but also applies to the demand for capacity at the system level. That implies modifications to the existing system demand curve may be needed, in order to ensure the system curve and zonal curves work together properly.

In sum, the marginal reliability benefit function in (3) is the basis for how the ISO proposes to determine capacity demand curves for the FCA. Before discussing some additional properties of these curves, it is useful to first explain how to extend this framework for a constrained system with capacity zones. This we address next.

Demand Curve for an Import Capacity Zone

The previous section presented the engineering-economic foundations for capacity demand curves, and for simplicity used formulas applicable to an unconstrained system without capacity zones. In this section, we explain how to extend this framework to derive zonal demand curves for an import-constrained capacity zone.

■ **EENS In a System With Zones.** As noted previously, the expected energy not served depends on both the total capacity in the system, and the amount of capacity located in any constrained zones. The precise way zonal constraints affect EENS is important to the derivation of zonal demand curves, so we step through that first.

Conceptually, the demand for capacity in an import zone – above and beyond the demand for capacity at the system level – should be based on the *additional* EENS in the system that arises due to the existence of a zonal import limit. For simplicity, assume there are two capacity zones in the system:

[6] These inputs are summarized in ISO's annual *ICR Related Values Report,* available at http://www.iso-ne.com/system-planning/resource-planning/installed-capacity-requirements. Preliminary values are presented and discussed with stakeholders at the PSPC during the year preceding each auction; for FCA10, updated ICR related values are available in the August 2015 PSPC Meeting materials at http://www.iso-ne.com/committees/reliability/power-supply-planning.

An import-constrained zone ('ICZ'), and the rest-of-system ('ROS') zone. In general, the total EENS in the system can be decomposed into the sum of two components: (1) The EENS in an *unconstrained* system (*i.e.*, without any zones) with aggregate capacity level Q_{SYS}, and (2) the *additional* EENS due to the zonal import limit, when the capacity level in the import-constrained zone is Q_{ICZ}. Stated as a formula,

$$\underbrace{EENS(Q_{ICZ}, Q_{SYS})}_{\text{Total EENS}} = \underbrace{EENS_{SYS}(Q_{SYS})}_{\text{Unconstrained System EENS}} + \underbrace{EENS_A(Q_{ICZ}, Q_{SYS})}_{\text{Additional EENS due to Zonal Import Limit}} \qquad (5)$$

The left-hand side of this equation is the total expected energy not served in this constrained system when there is Q_{ICZ} capacity in the import-zone, and Q_{SYS} capacity in the system. On the right-hand side, the first term is the EENS system-wide if, counter to fact, there was no zonal import limit constraint and the system has Q_{SYS} capacity. (Since this term is not dependent on Q_{ICZ}, we use the notation $EENS_{SYS}$ for this term in (5) to indicate it is a different mathematical function).

The second term on the right-hand side of (5) has an important and different interpretation. Mathematically, this term represents the *additional* EENS in this two-zone constrained system, above and beyond the EENS that would occur in an unconstrained system with the same total capacity Q_{SYS}, if there is only Q_{ICZ} capacity in the import zone. (We use the subscript 'A' in the last term in (5) to indicate this is the additional EENS due to the import limit, as distinct from both the total EENS and the unconstrained system $EENS_{SYS}$.)

■ **Marginal Reliability Impacts.** As before, we will derive the capacity demand curve based, in part, on the marginal reliability impact of capacity in the zone. For zones, however, the marginal reliability impact of capacity is a more subtle concept than it is for the system as a whole. This is because there are two different ways to measure of the marginal reliability impact of capacity in an import zone:

1. *Marginal Reliability Impact of Capacity Substitution.* This is the reduction in total EENS if 1 MW of capacity is substituted (or shifted) out of the ROS zone and into the ICZ, holding total system capacity constant.

2. *Marginal Reliability Impact of Capacity Additions.* This is the reduction in total EENS if 1 MW of capacity is added in the ICZ, and the ROS zone capacity is held constant (so total system capacity also increases by 1 MW).

In principle, a demand curve can be derived for a capacity zone based on either of these two marginal reliability impact measures. But they are different things. We use the former, because both the FCA's clearing algorithm and the interpretation of the zonal demand curves are simplified using the former instead of the latter. In particular, by using the first measure of zonal marginal reliability impacts, the zonal demand curves (under the proposed method) will properly calculate *congestion prices* between capacity zones.

■ **Evaluating Zonal Marginal Reliability Impacts.** At this point, a few additional elements will lead us to a simple way to interpret (and evaluate) zonal capacity demand curves. First, to obtain the zonal capacity demand curve, we determine the marginal reliability impact of capacity substitution. This equals the gradient of the 'additional EENS' function in equation (5). Specifically, if there is Q_{ICZ} capacity currently in the ICZ, then the marginal reliability impact of substituting one unit of capacity out of the ROS zone and into the ICZ (*i.e.*, while holding total system capacity constant at Q_{SYS}) is the partial derivative:

$$\frac{\partial}{\partial Q_{ICZ}} EENS_A(Q_{ICZ}, Q_{SYS})\Big|_{Q_{SYS}} \qquad (6)$$

This looks complicated, but stay with it for just another moment – there's something important just around the curve (so to speak). Although this is a measure of the marginal reliability impact, it isn't practical (in this form) for a demand curve – the expression in (6) is a function of two variables, Q_{ICZ} and Q_{SYS}. To get a zonal demand curve that is a function of one variable, Q_{ICZ}, we can evaluate this marginal reliability impact when the system's capacity is at the equilibrium level consistent with the Reliability principle (*viz.*, the '1-in-10' capacity level in an unconstrained system). That is, we can evaluate the zonal marginal reliability impact of capacity substitution as:[7]

$$\textit{Zonal Marginal Reliability Impact} = \frac{\partial}{\partial Q_{ICZ}} EENS_A(Q_{ICZ}, Q_{SYS})\Big|_{Q_{SYS}=1:10} \qquad (7)$$

This is a negative, smoothly increasing function of the amount of capacity in the import zone. That means substituting additional capacity into an import zone reduces EENS, but has a progressively a diminishing marginal impact.

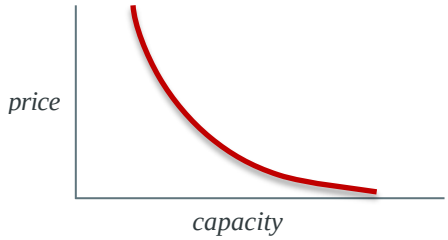
Importantly, the ISO's reliability planning models can calculate all of the relevant zonal EENS values, and their derivatives, presently. For an import zone, this is a two-step process: First, the unconstrained system EENS (and 1-in-10 capacity level) are determined using the existing reliability planning model assuming no import constrained zones, which yields the first term in (5). Second, the same model is run after imposing the zonal import limits to obtain the second term in (5), and its gradient in (7). Neither of these calculations involves any additional information beyond that presently used to calculate the various ICR-related values for each FCA.

■ **Demand Curve for an Import Zone.** The demand curve for an import capacity zone is given by its marginal reliability benefit function, in a manner consistent with the engineering-economic foundations described earlier. Using the notation P_{ICZ} to denote the price specified by the import zone demand curve at Q_{ICZ} capacity, we have:

$$\textit{Zonal Demand:} \qquad P_{ICZ}(Q_{ICZ}) = -PF \times \textit{Zonal Marginal Reliability Impact}$$

[7] Using Q_{SYS} at the 1:10 (NICR) level in (7), as opposed to a different level, is not crucial; the marginal reliability impact curve may not vary materially for different levels of Q_{SYS} over the empirically-relevant range of system capacity levels.

A demand curve for an import capacity zone will have this general shape:



This curve's shape is intuitive: It decreases with additional capacity, and gets progressively flatter as more capacity is added. The reason for this shape is simple: When there is relatively little capacity in a zone, there are many expected hours with lost load (in the zone), and any further reduction in capacity has a large, adverse marginal reliability impact. In other words, when there is relatively little capacity, the curve must rise steeply as capacity falls. In contrast, when there is ample capacity, there are few expected hours with lost load (in the zone), and any additional capacity in the zone may be rarely used. Thus, at higher capacity levels in the zone, the marginal reliability impact does not change much with additional capacity so the curve declines gradually.

In sum, to respect these fundamental (physical) properties of how capacity affects expected energy not served, the demand curves must have the 'scooped' shape shown in the figure above.

■ **Zonal Demand Curves Specify Congestion Prices.** The zonal demand curve derived above has a very important interpretation. Because it is based on the *additional* EENS that occurs due to the zonal import limit, the prices specified by the zonal demand curve represent the *additional* amount that should be paid for capacity in the zone – that is, in addition to the system capacity clearing price. In other words, the zonal demand curve is a *congestion pricing* curve.

An analogy may help. Recall the basic interpretation of congestion prices in the energy market, which arise whenever a constraint limits the amount of energy that can be imported into a particular location. Ignoring energy losses (which are not relevant here), the locational and the system prices differ by the congestion price across the constraint:

Locational Marginal Price = System Energy Price + Locational Congestion Price

An import-zone capacity demand curve, as derived above, works on the same logic:

Zonal Capacity Price = System Capacity Price + Zonal Congestion Price

Consider the following example. Suppose the system demand curve clears at a price of $6 per kw-month. Imagine that, at the quantity of capacity cleared in the import zone, the zonal demand curve specifies a price of $2 per kw-month. This $2 price is not the total price paid to capacity in the zone, it is the *congestion price* for capacity in the zone. The total price paid to capacity in the zone is given by the system's capacity clearing price of $6 *plus* the zonal congestion price of $2, or $8 per kw-month in total.

Continuing the example, imagine that a particular capacity resource offers into the market at $7 per kw-month. If it is located in the import capacity zone, it would clear and be paid the zonal clearing price of $8 per kw-month. If another resource also offers $7 per kw-month and is located in the ROS zone, it would not clear the auction (because its offer price exceeds the system capacity clearing price of $6 per kw-month).

■ **Pricing Congestion with a Zonal Capacity Demand Curve Makes Sense.** It is natural wonder why interpreting the zonal demand curve in terms of congestion pricing make economic sense. This is a consequence of how the import-zone capacity demand curve is being evaluated. Specifically, the zonal marginal reliability impact (in expression (7)) is the change in EENS if we *substitute* (or shift) one unit of capacity out of the ROS zone and into the ICZ. Substituting capacity from the ROS zone into an ICZ has a reliability benefit, in general, because it helps reduce EENS inside the zone when the zonal import constraint is binding and incremental capacity in the ROS would not.

In other words, a zonal demand curve based on the marginal reliability impact of capacity substitution tells us how much more 1 MW of additional capacity is worth if procured in the ICZ, *instead of* being procured in the ROS zone. That is, it indicates what the price *difference* should be between the ICZ and the ROS zone. This price difference across a constrained interface is, of course, what we normally call the congestion price.

■ **Import Limits and LSR.** It is important to note that no-where in the derivation of the zonal capacity demand curve is the current Local Sourcing Requirement imposed as a 'fixed' requirement. The zonal demand curve does account, in the design directly, for the import interface transfer limit into the zone. This is because the marginal reliability impact of zonal capacity is derived from the *additional* EENS due to the zonal interface import limit (*viz.*, the last term in expression (5)). With additional import transfer capability, the zonal demand curve becomes flatter; and if there was no import interface limit at all, the last term in (5) would be zero and the zonal demand curve would be a flat line at zero. Put differently, a constraint that never binds always has a congestion price of zero.

■ **Marginal Reliability Benefit of Capacity Additions.** It is useful to connect a few more dots between the system and zonal prices. As just explained, the price specified by the zonal capacity demand curve is the zone's capacity congestion price, because it is based on the reliability impact of *substituting* capacity between the ICZ and the ROS zone. Let's now answer: What is the marginal reliability benefit of adding capacity inside the import-constrained zone, if the MW is a net addition to the system (rather than being substituted between zones)?

The answer is the *sum* of the system price and the zonal congestion price. In other words, the total price to be paid to capacity in an import constrained zones is the system price plus the zonal congestion price, like congestion pricing normally works.

To see how this holds in this context, one can think of our net capacity addition inside the ICZ in two conceptual steps: A reliability benefit from adding another MW of capacity into the ROS zone, *plus* an additional potential reliability benefit from shifting that MW out of the ROS and into the ICZ. The marginal reliability benefit of the first step is given by the system demand curve (as explained in the prior section on economic foundations), and yields the system capacity clearing price. The marginal reliability benefit of the second step, substituting the new MW from the ROS into the ICZ, is given by the zonal capacity demand curve that specifies the zonal congestion price. The total effect of adding

an incremental MW of capacity inside the import-constrained zone, if it is a net addition to the system, is therefore the sum of the two prices. In other words, paying all capacity inside the import-constrained zone the system clearing price *plus* the zonal demand curve's congestion price will, in fact, compensate capacity investment inside the ICZ commensurate with its marginal reliability benefit.

Demand Curve for an Export Capacity Zone (New)

An important feature of this engineering-economic approach to developing capacity demand curves is that also provides a logical basis for defining demand curves for an export-constrained capacity zone. The approach is analogous to that for import zones, with various signs reversed. We step through the analysis here.

■ **EENS With Export Zones.** As with an import zone, the demand for capacity in an export zone – above and beyond the demand for capacity at the system level – is based on the *additional* EENS in the system that arises due to the existence of a zonal export limit. Here, the additional EENS will generally be positive – that is, the export limit *increases* the expected energy not served.

Assume now there are three capacity zones in the system: An import-constrained zone (ICZ), an export-constrained zone (ECZ) and the rest-of-system (ROS) zone. The total EENS in the system can be decomposed into the sum of three components: (1) The EENS in an unconstrained system (*i.e.*, without any zones), (2) the additional EENS due to the import zone transfer limit, and (3) the additional EENS due to the export zonal transfer limit. Stated as a formula,

$$\underbrace{EENS(Q_{ICZ}, Q_{ECZ}, Q_{SYS})}_{\text{Total EENS}} = \underbrace{EENS_{SYS}(Q_{SYS})}_{\text{Unconstrained System EENS}} + \underbrace{EENS_{AI}(Q_{ICZ}, Q_{SYS})}_{\text{Additional EENS due to Zonal Import Limit}} + \underbrace{EENS_{AE}(Q_{ECZ}, Q_{SYS})}_{\text{Additional EENS due to Zonal Export Limit}} \quad (8)$$

The left-hand side of this equation is the total expected energy not served in this constrained system when there is Q_{ICZ} capacity in the import zone, Q_{ECZ} capacity in the export zone, and Q_{SYS} capacity in the system. On the right-hand side, the first two terms are the same as discussed previously for an import zone (see equation (5) *ff.*). The new, final term is the additional EENS in this three-zone constrained system if there is Q_{ECZ} capacity in the import zone. Note the last two terms do not depend on the other constrained zone's capacity directly, because the additional EENS in an export zone is not impacted by the import-zone limit, and vice versa. (We use subscripts '*AI*' and '*AE*' in the two last terms in (8) to indicate these additional EENS functions are distinct from one another, and from the unconstrained system $EENS_{SYS}$.)

■ **Evaluating Zonal Marginal Reliability Impacts.** Proceeding similarly to the case with an import zone, we evaluate an export zone's marginal reliability impact as the change in the 'additional' EENS function applicable to the export zone (*i.e.*, the gradient of the last term in expression (8)):

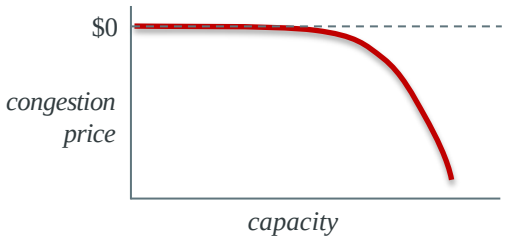
$$\textit{Zonal Marginal Reliability Impact} = \frac{\partial}{\partial Q_{ECZ}} EENS_{AE}(Q_{ECZ}, Q_{SYS})\Big|_{Q_{SYS}=1:10} \qquad (9)$$

As with an import zone, this is the marginal reliability impact of capacity substitution from the ROS zone into the ECZ. This is a positive, smoothly increasing function of the amount of capacity in the export zone. That means, after a certain level, substituting additional capacity into an export zone *increases* overall EENS, and has a progressively greater marginal reliability impact.

■ **Demand Curve for an Export Zone.** The demand curve for an export capacity zone is given by its marginal reliability benefit. Using the notation P_{ECZ} to denote the price specified by the export zone demand curve at Q_{ECZ} capacity, the demand curve is determined by the same formula as for an import zone:

$$\textit{Zonal Demand:} \qquad P_{ECZ}(Q_{ECZ}) = -PF \times \textit{Zonal Marginal Reliability Impact}$$

As with an import zone, this demand curve represents a congestion price. With an export zone, however, this price is always negative or zero; that is, resources in an export-constrained zone are paid the same, or less than, resources in the rest-of-system. Graphically, the demand curve for an export capacity zone will have this general shape:



This curve's shape is intuitive: It decreases with additional capacity. When there is relatively little capacity in a zone, the export transmission limit rarely (or never) binds, so the marginal reliability impact of capacity substitution into the zone is negligible and the curve is flat. However, when there is a high level of capacity in the export zone, there are many expected hours when the export limit is binding and additional capacity in the export zone may be rarely used. Thus, at higher capacity levels in the export zone, substituting more capacity from the ROS into the ECZ ('behind' the constraint) has a progressively adverse marginal reliability impact. The curve therefore begins to fall steeply at higher levels of capacity in an export zone.

■ **Export Zone Demand Curves Specify Congestion Prices.** Since the zonal demand curve derived above is a *congestion pricing* curve, it does not directly determine the total price paid to resources in the zone. Rather, as with other zones, they are paid based on the sum of the system price and the zonal congestion price:

$$\textit{Zonal Capacity Price} = \textit{System Capacity Price} + \textit{Zonal Congestion Price}$$

An example may help. Suppose the system demand curve clears at a price of $6 per kw-month. Imagine that, at the quantity of capacity cleared in the export zone, the export zone demand curve

specifies a congestion price of $ –2 per kw-month. The total price paid to capacity in the zone is given by the system's capacity clearing price of $6 plus the zonal congestion price of $–2, or $4 per kw-month in total.

Continuing the example, imagine that a particular capacity resource offers into the market at $5 per kw-month. If it is located in the export capacity zone, it would not clear since its offer exceeds the zonal capacity clearing price of $4. If another resource also offers $5 per kw-month and is located in the ROS zone, it would clear the auction because its offer price is less than the system capacity clearing price of $6 per kw-month.

Importantly, since capacity supply offers/bids are non-negative, an export zone's congestion price cannot be so large (in magnitude) as to produce a negative zonal capacity clearing price paid to capacity providers. That is, if the system clearing price is $6 per kw-month as before, the most negative hypothetical congestion price that could occur across the export zone interface would be $–6 per kw-month. As a result, the lowest possible total price paid to capacity a zone is zero.

■ **Interpreting the Export Demand Curve.** As with an import zone, an export zone demand curve is based on the marginal reliability impact of capacity substitution. It tells us how much less 1 MW of additional capacity is worth if procured in the ECZ, *instead of* being procured in the ROS zone. That is, it indicates what the price *difference* should be between the ECZ and the ROS zone. Export-zone congestion prices are negative (or zero) because incremental capacity in the ECZ is an imperfect substitute for capacity in the ROS: incremental capacity in the ROS helps reduce EENS in the ROS when the zonal export constraint is binding, but incremental capacity in the ECZ would not.

■ **Export Limits and MCL.** It is important to note that no-where in the derivation of the zonal capacity demand curve is the current Maximum Capacity Limit rating imposed as a 'fixed' limit. The export zonal demand curve does account, in the design directly, for the export interface transfer limit out of the zone. This is because the marginal reliability impact of zonal capacity is derived from the *additional* EENS due to the zonal interface export limit (*viz.*, the last term in expression (8)).

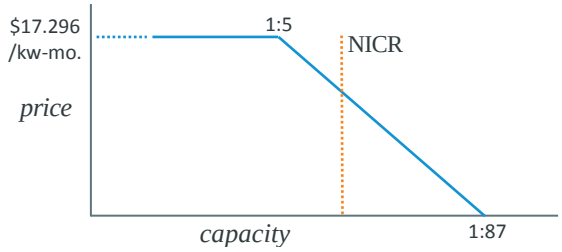
With additional export transfer capability, the export zonal demand curve becomes flatter; and if there was no export interface limit at all, the last term in (8) would be zero and the zonal demand curve would be a flat line at zero. As with any transfer limit, a constraint that never binds always has a congestion price of zero.

■ **Practical Implications.** There's one caveat to this simple and intuitive economic logic of demand curves and congestion pricing. Its foundations also imply that the *system* demand curve, and therefore the system capacity clearing price, should also determined on the basis of the marginal reliability benefit analysis. In other words, the system and zonal demand curves should be viewed as an integrated system for procuring capacity, and designed cognizant of one another (*viz.*, on the basis of (8)); if they are not, the curves may procure too much capacity in one zone and too little in another (relative to the actual marginal reliability impact in each zone). This brings us to why certain modifications to the system sloped demand curve are desirable, and that we will discuss presently.

Concerns with the Existing System Demand Curve (New)

Based on its marginal reliability analyses for capacity demand curves, the ISO recommends the region make certain modifications to the system-level capacity demand curve, in addition to adopting zonal curves. In this section, we summarize the main concerns we have identified with the existing demand curve. We explain the specific proposed modification to address these concerns in the subsequent section.

Presently, the FCA uses a linear system-level demand curve. Graphically, it looks like this:



As a preliminary observation, this linear demand function resulted from a general consensus among stakeholders and the ISO that it represents an improvement upon the prior practice (in FCA 8 and prior auctions) of using a fixed requirement, or vertical capacity demand curve, at NICR. Indeed, the existing sloped system demand curve is a considerable improvement in that respect.

With the benefit of additional time and analysis since the current system demand curve was adopted, we have identified three concerns with this linear system demand curve.

■ **Current system curve poorly reflects capacity's marginal reliability impact.** The existing system curve does not account for the fact that, from a reliability perspective, the change in the marginal reliability impact with an additional unit of capacity is very different when the system is much *below* NICR than when it is *above* NICR.

When system capacity is below NICR, the change in the marginal reliability impact if the FCA clears one less increment of capacity is quite high. At capacity levels approaching the 1-in-5 reliability level (*i.e.*, LOLE = .2, or 33,076 MW for FCA10), the ISO's reliability planning models indicate the EENS exceeds 1600 MWh per year and increases rapidly as capacity levels decline further. The system demand curve should therefore rise more steeply – to ensure the FCA clears closer to NICR – as capacity levels fall below NICR.

In contrast, when the system capacity is well above NICR, the change in the marginal reliability impact with each increment of capacity is quite low. For example, at the 1-in-50 capacity level (*i.e.*, LOLE = .02, or approximately 36,400 MW), the EENS is approximately 60 MWh annually and decreases very gradually, and at a progressively slower rate, with additional capacity. In other words, to reflect the marginal reliability impact properly, the system demand curve should become steadily 'flatter' as it decreases.

Viewed in terms of marginal reliability impacts more directly, the ISO's reliability planning models

indicate that the change in the system's marginal reliability impact with another increment of capacity is approximately 10 times as large at the 1-in-5 capacity level as it is at the 1-in-50 level. This means the system demand curve should have approximately *10 times the slope* at 1-in-5 as it does at 1-in-50, if it is to properly reflect how incremental capacity affects reliability. The existing linear demand curve, which has the *same* slope at both 1-in-5 and 1-in-50, was not selected with this pertinent information at hand. As a result, it poorly reflects how reliability changes when capacity levels differ from the 1-in-10 level.

There are economic consequences to this mismatch between the marginal reliability impacts of capacity investment and the prices resources would be paid (as specified by the demand curves). Specifically, the existing linear demand curve may tend to undercompensate resources for the reliability benefits they provide – and send too low a price signal for new investment – when the system is significantly below NICR, because the demand curve should be steeper in this region than it is today. The existing linear demand curve will also tend to overcompensate resources for the reliability benefits they provide – and send too high a price signal for new investment – over a broad range of capacity levels whenever the system is above NICR, because the demand curve should be much lower, and much flatter, in that region than it is today.

■ **Current system curve procures more capacity than necessary to meet the reliability objective (Principle 1).** The second concern relates to where the system curve is positioned relative to the intersection of NICR and Net CONE. With the benefit of additional analysis that was not available at the time the linear system demand curve was selected, it appears the existing system curve will tend to procure more capacity than necessary to meet the 1-in-10 (LOLE = 0.1) objective, on average over time.

This concern is primarily a consequence of the fact that the system demand curve should be convex, on the basis of engineering-economic considerations discussed previously. With a convex demand curve, the same average reliability level (*i.e.,* LOLE = .1) can be achieved with a lower average capacity and lower average cost.

A simple (hypothetical) example may help. Assume the FCA clears each year at either Net CONE minus $2 per kw-month, Net CONE, or Net CONE plus $2 per kw-month, with all three outcomes equally likely. If a system demand curve has the same slope as the current linear system demand curve, then to achieve an average LOLE of 0.1 over time (using current New England system parameters) in this simple example requires capacity levels of 34,662 MW, 34,203 MW, and 33,743 MW at each price (respectively). The capacity procured is 34,203 MW on average (over time).

Now imagine we use a convex demand curve. This enables different quantities to be cleared at each price while achieving the same average LOLE. Specifically, if a convex demand curve procures the capacity levels of 34,418 MW, 34,151, and 33,925 MW at each price (respectively), the average LOLE is still 0.1 (using current New England system parameters). The average capacity procured would be lower, however at 34,165 MW. Because the average capacity procured is lower, average total costs are also lower: In this (hypothetical) example, average capacity costs are $1.5 million less with the convex curve than with the linear curve, while achieving the same average reliability.

This phenomenon isn't terribly intuitive, but it is important. It occurs because LOLE is a (highly) non-linear, convex function of capacity. Using a convex demand curve is much better able to match

thus fundamentally non-linear relationship, enabling the FCA to achieve the 1-in-10 target *on average* with a lower level of installed capacity on average.

The inefficiencies that arise using a linear demand curve in this fundamentally non-linear setting are the reason that, when the current linear system demand curve was developed, it was necessary to 'tune' the linear curve so it passes considerably to the 'right' of NICR at the Net Cone value. With a convex curve, it is not necessary for the demand curve to pass similarly far to the 'right' of NICR at Net Cone, resulting in lower average capacity procurement – and a lower average total cost of capacity – to deliver the same average reliability. While the difference in total cost in this intentionally-simplified example are modest (*i.e.*, only $1.5 million), with other supply offers and larger potential variation in clearing prices over time, the efficiency consequences may be considerably larger.

The broad point is there are pure efficiency gains to be realized when a curve's shape reflects sound engineering-economic fundamentals. The region can achieve same reliability target on average, at a lower average cost to society, by modifying the existing system curve to reflect these fundamentals using the MRI information now available.

■ **Current system curve does not cost-effectively allocate capacity procurement across capacity zones.** The third concern relates to the allocation of capacity purchases across zones. When zonal demand curves are employed, rather than fixed zonal requirements, the same system overall reliability can often be maintained by procuring a little more capacity in one zone, and a little less capacity in another (but not at a 1-for-1 rate, in general). It is therefore necessary to have a principle to decide how much to demand in each zone when there are multiple choices that achieve the same overall reliability level.

In this context, it becomes important to select system and zonal demand curves that are *cost-effective.* In simple terms, this means that when there are multiple choices regarding how much capacity to demand in each zone (including the Rest-of-System zone) that would achieve the same overall reliability, the quantities to be chosen should minimize the total bid-cost of capacity overall.

The ISO's concern here is that, based on the additional analysis now available, the existing system demand curve does not satisfy this basic cost-effectiveness design principle. Specifically, when prices separate, there is generally a different level of system capacity than that specified by the existing system demand curve that would achieve (a) the same or better system reliability (b) at lower total bid-cost.

This cost-effectiveness principle between zones was not considered when the existing system demand curve was developed, in part because the existing system demand curve was developed in isolation of zonal demand curve considerations. The system demand curve and the zonal demand curves should not be considered in isolation, however, as they jointly determine – in combination with the capacity supply bids/offers – how much capacity will is cleared in each zone (including the Rest-of-System capacity zone).

The concept and logic of cost-effective demand curves when there are both system-level and zonal demand curves is addressed in detail presently. There we explain why the status quo is not cost-effective, and why cost-effectiveness requires the system and zonal curves to be developed as an

integrated process. Before addressing cost-effectiveness in greater detail, however, it will be useful to summarize the relevant modifications the ISO recommends to the system-level demand curve.

System Demand Curve Modifications

To address all of the foregoing concerns with the existing system demand curve, the ISO proposes to modify it in conjunction with the adoption of zonal demand curves. This section summarizes the proposed system demand curve modifications.

The ISO's proposed modifications to the system demand curve are based on the same foundations discussed earlier in this memorandum (see pages 3-5). Using the notation P_{SYS} to denote the price specified by the system demand curve at Q_{SYS} capacity, the modified system demand curve would be determined as:

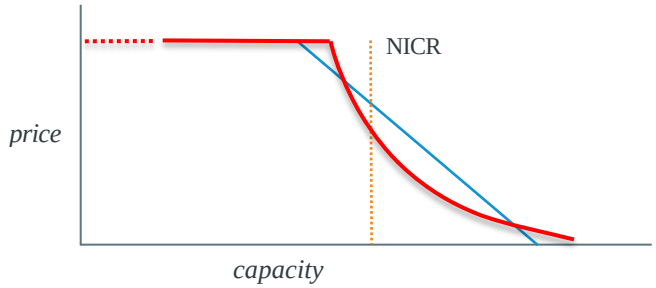
$$\textit{System Demand:} \qquad P_{SYS}(Q_{SYS}) = -PF \times \textit{Marginal Reliability Impact}$$

Here, the marginal reliability impact is determined by the change in EENS in an unconstrained system with respect to system capacity (*see* expression (3) on page 5):

$$\textit{Marginal Reliability Impact} = \frac{d}{dQ_{SYS}} EENS_{SYS}(Q_{SYS})$$

Note that the system-level MRI of capacity is always calculated in the same way, regardless of the number and types of capacity zones in the system. This ensures that the assessment of the marginal reliability impact of adding system capacity is robust to the zonal configuration.[8]

Since a picture is worth a thousand words, we expect the appropriately-modified system demand curve should look generally like this. The new (red) curve has the 'scooped', or convex, shape and the existing (blue) curve is the downward sloping straight line:



[8] Mechanically, this can be verified by noting that the system-level MRI of capacity is the same whether calculated using the first-term on the right hand side of expression (5) (on page 8, in the case with only import zones), or if calculated using expression (8) (on page 12, in the case with both import and export constrained zones). In each case, and consistent with the current process for determining the NICR, the system-level demand curve and system-level MRI is determined for an unconstrained system.

The flat section of both curves is the existing price cap (the FCA Starting Price), which we do not propose to modify. As with any capacity demand curve that reflects the marginal reliability impact of capacity, this modified system curve gets progressively flatter with additional capacity, reflecting the rapidly diminishing marginal reliability benefit of additional capacity over the relevant range of possible system capacity levels.

Although the existing system demand curve was developed in isolation of any specific zonal demand curve design, it is important to note that the system curve affects the allocation of capacity purchases among zones – in particular, both the system and the zonal curves determine demand in the ROS zone. This brings us back to the topic of cost-effective procurement between capacity zones, next.

Cost-Effectiveness Between Capacity Zones (New)

One of the central design principles guiding the ISO's proposal is to procure capacity cost-effectively among zones. In this section, we explain this principle and illustrate it using simple examples. We then explain why the ISO's proposed method achieves this principle, and why other alternative zonal demand approaches (including the status quo) do not.

■ **Concepts.** First, some clarifying terminology will be useful. We say a (set of) demand curves is *cost-effective* if, for each possible clearing price in each zone, there is no way to modify the cleared capacity levels among zones that achieves (a) the same or better system reliability at (b) lower total bid-cost. Stated in other terms, if a set of demand curves is <u>not</u> cost effective, then different demand curves would deliver the same reliability at lower total bid-costs.

By itself, cost-effectiveness would seem a difficult principle to argue with. In essence, it asks that any set of zonal demand curves pass a basic test: Can we alter the quantities demanded at the given prices, by potentially purchasing a little more in one zone and a little less in another, in a way that maintains the same overall system reliability but lowers total auction bid-cost? If so, then the demand curve values at those prices could be modified, with no loss in reliability, to lower cost.

Importantly, cost-effectiveness between zones is not assured by satisfying the reliability and sustainability design principles alone (Principles 1 and 2). Those principles apply to the average revenue and average system reliability achieved by the design (over time). Cost effectiveness is a much stronger principle. It asks that at *any* set of clearing prices that could occur in a given auction, the demand curves should specify quantities to purchase in each zone that minimize total bid-cost for the particular level of reliability achieved in that year's auction (which could be higher or lower than '1-in-10' in a given year).

Cost-effectiveness between zones is easiest to understand with the help of a few examples, which we provide next.

■ **Example.** To explain these ideas, we first consider a simple example. This example explains how to check whether a set of demand levels for a system with two capacity zones is cost effective or not. It also explains why.

First, assume a system with two zones: One export-constrained zone, and the ROS zone. In this hypothetical example, we assume the system clears at $9 per kw-month, and the quantity given by the system at this demand curve is 34,000 MW. We assume the ECZ clears lower, at $6 per kw-month, and at that price the quantity given by the ECZ demand curve is 9,500 MW. These values are collected in the table below:

Example 1a.	System	ECZ	Ratio	*Interpretation*
Total Price	$9	$6	**1.5**	*More costly in System*
Demand (MW)	34,000	9,500		
Total MRI (h/yr)	– 2/3	– 1/3	**2.0**	*More effective in System*

Now, to assess the reliability impacts of capacity in each zone, we require their marginal reliability impact values. Here, as shown in the table above, we assume that the MRI of additional capacity in the system is – 2/3 MWh; that means one incremental MW of capacity in the system would reduce the system's total EENS by 2/3 MWh per year. However, capacity in the ECZ is not as effective, with an MRI of capacity additions of only – 1/3 MWh.[9] The MRI of capacity additions in an ECZ is less than (or, in some cases, equal to) the MRI in the system, because a unit capacity in ROS can help avoid lost load in the ROS when the export interface limit is binding, but a unit of capacity in the ECZ cannot.

The most important part of the table above is the ratios column. They reveal that, at the margin:

- Additional capacity is 50% more costly in the ROS than in the ECZ ($9 v. $6);
- Additional capacity is *twice* as effective in the ROS than in the ECZ (– 2/3 v. – 1/3).

This implies the demand curves that generated these quantities are *not* cost effective. They demand too much capacity in the ECZ relative to the ROS. Reducing demand in the ECZ, and increasing it in the ROS – at the right rates – can deliver the same reliability at lower total cost.

To see this, let's consider what happens if we procure 1 MW *less* capacity in the ECZ. First, note the impact of lowering the ECZ demand on EENS: – 1 MW × (– 1/3 MRI) = + 1/3 MWh increase in EENS. If we procure a little *more* capacity in the ROS, however, we can offset this reliability impact. Crucially, we do not need to buy quite as much capacity in the ROS, because it is more effective there. Specifically, if we demand an additional .5 MW of capacity in the ROS, the impact of this higher procurement on EENS is + .5 MW × (– 2/3 MRI) = – 1/3 MWh.

[9] In Example 1a, we list the MRI of capacity *additions* in each zone. An underlying assumption is that the MRI of capacity *substitution* from the ROS into the ECZ is +1/3 MWh/yr (an export zone's MRI of substitution is always positive). In general, the MRI of capacity additions in any zone is equal to the system MRI plus the MRI of capacity substitution. Applied in this example, the MRI of capacity additions in the ECZ is – 2/3 MWh/yr + 1/3 MWh/yr = – 1/3 MWh/yr, as shown in the table.

Put this in more general terms, incremental capacity in the ROS and in the ECZ are *partial substitutes*. Both help improve reliability, but not equally so. Buying just a little more of the one that is more effective (*i.e.*, has the higher absolute MRI), and much less of the one that is less effective, can achieve the same overall level of reliability.

Now, back to cost. Let's collect our numbers for the modified demands in a new table, below. In the first row, we indicate the change we are checking to test the cost-effectiveness of the original demand curves: Buying 1 MW less in the ECZ, and .5 MW more in the ROS. The second row shows the change in EENS due to each demand change. And the third row shows the change in cost.

Example 1b.	**ROS**	**ECZ**	**Total**	***Interpretation***
Δ Capacity	+ .5 MW	– 1 MW	**– .5 MW**	*Less total capacity*
Δ EENS (MWh)	– 1/3	+ 1/3	**0**	*Same reliability*
Δ Cost ($/mo)	+ $4,500	– $6,000	**– $1,500**	*Lower total cost*

By demanding 1 MW less in the ECZ at a price of $6/kw-month, cost falls by $6,000 per month. Demanding .5 MW more in the ROS raises cost, but not by as much: .5 MW × $9 / kw-mo. × 1000 = $4,500 per month. In sum, the results in the table above show that the original demand curves in this example are not cost effective: Demanding different quantities can achieve the same overall reliability, at a lower total cost.

Stepping back a bit, this simple example illustrates a core issue for demand curve design with capacity zones. When zonal demand curves are employed, rather than fixed zonal requirements, the same overall reliability can often be achieved by procuring a little more capacity in one zone, and a little less capacity in another (but not at a 1-for-1 rate, as this example illustrates). It is therefore necessary to decide how much to demand in each zone, since there are multiple choices that achieve the same overall reliability level. The ISO's proposed design principle to resolve this decision is the cost-effectiveness criterion. This means that when there are multiple choices regarding how much capacity to demand in each zone (including the Rest-of-System zone) that would achieve the same overall reliability, the quantities demanded should minimize the total bid-cost of capacity overall.

■ **Status Quo is Not a Cost Effective Solution.** One of the ISO's concerns is that, based on the additional insights and MRI information now available, the FCA's status quo – with the current system demand curve and fixed zonal requirements – is not cost-effective.

Here's a simple way to see this, extending the prior example. Here we use actual quantities specified by the current linear system demand curve, and a (non-hypothetical) export-constrained zone corresponding to the Northern New England (NNE) export zone tested for FCA10. The indicative maximum capacity limit for the NNE zone is 8,830 MW.[10]

[10] See the ISO's November 16, 2015 presentation at http://www.iso-ne.com/static-assets/documents/2015/11/a2_fcm_zonal_development_3_review_of_determinations_for_fca_10.pdf

Assume as before that the clearing prices in the system and NNE export zone are $9 and $6, respectively (this analysis can be performed similarly with any other prices). The quantities specified at these prices by the current system demand curve and by a 'vertical' zonal demand curve at the NNE maximum capacity limit are shown below. Using the ISO's marginal reliability analysis results (for FCA 10 inputs), we have calculated the total MRI of capacity additions in each zone at these two capacity levels.[11] These are provided in the following table:

Example 2a.	System	ECZ	Ratio	*Interpretation*
Total Price	$9	$6	**1.5**	*More costly in System*
Demand (MW)	34,983	8,830		
Total MRI (h/yr)	– .312	– .233	**1.34**	*Relatively less effective*

The ratio column in this table indicate that, at these prices, capacity is 50% more costly in the ROS than in the ECZ ($9 v. $6), but capacity is only 34% more effective in the ROS than in the ECZ, at the margin. This implies the quantities procured at these prices are *not* cost effective. Since the price ratio exceeds the effectiveness ratio, these quantities procure too much capacity in the ROS relative to the ECZ. (Note, this is the opposite situation from Example 1, where there was too much capacity in the ECZ relative to the ROS).

Reducing demand in the ROS, and increasing it in the ECZ – at the right rates – can deliver the same reliability at lower total cost. To see this, consider procuring 1 MW more capacity in the ECZ, and procuring .75 MW (= 1/1.34) less in the ROS. The change in capacity levels, EENS, and cost is summarized in the following table:

Example 2b.	ROS	ECZ	Total	*Interpretation*
Δ Capacity	– .75 MW	+ 1 MW	**– .25 MW**	*Less total capacity*
Δ EENS (MWh)	+ .233	– .233	**0**	*Same reliability*
Δ Total Cost ($/mo)	– $6,750	+ $6,000	**– $750**	*Lower total cost*

In sum, there is a net efficiency gain from procuring different quantities than specified under the status quo, at least at these prices. Although we omit it here, similar analyses using other prices yield the same general finding: The status quo was not designed to procure capacity levels in different zones cost-effectively. Modifying the demand curves with this objective as a design principle has a clear and simple benefit – at any price levels, it can achieve the same reliability at lower total cost.

[11] At these capacity levels, the system MRI = –.312 MWh/yr and the NNE MRI of capacity substitution is +.079 MWh/yr, so the total MRI of capacity additions in the NNE zone is –.312 + .079 = –.233 MWh/yr, as shown in the table. For the underlying data, see http://www.iso-ne.com/static-assets/documents/2015/12/a09_iso_indicative_demand_curve_values_fca10_zones_12_03_15.xlsx .

■ **Procuring Capacity Cost-Effectively: Example.** By design, the ISO's proposed method for specifying demand curves (both for the system and each zone) solves this problem and procures capacity cost-effectively. We show this first with another example, which illustrates a general result. We then use the example to explain why the ISO's proposed method works generally, and why other demand curves approaches would not be cost effective.

Consider again the same two zones, the ROS and the export-constrained zone corresponding to the Northern New England (NNE) zone tested for FCA10. Building on the prior examples, assume the same prices as before (the same conclusion would hold at any other prices, for reasons explained further below). However, now consider the quantities demanded under the NNE zonal demand curve and the modified system demand curve under the ISO's proposal. These quantities, along with the corresponding MRI values, are shown in the table below.[12]

Example 3a.	**System**	**ECZ**	**Ratio**	***Interpretation***
Total Price	$9	$6	**1.5**	*More costly in System*
Demand (MW)	34,390	9,000		
Total MRI (h/yr)	– .498	– .334	**1.5**	*Same relative effectiveness*

The ratio column in this table indicates that, at these prices, capacity is 50% more costly in the ROS than in the ECZ ($9 v. $6), and simultaneously capacity is 50% more effective in the ROS than in the ECZ, at the margin. As in the previous examples, one could again consider procuring 1 MW more in the ECZ, and 2/3 MW (= 1/1.5) less in the ROS, which will again hold reliability (total EENS) unchanged. But in contrast to the previous examples, there would be no efficiency gain in doing so, as there is no further cost saving to be had by substituting incremental capacity between the zones.

Put simply, at the quantities demanded under the ISO's proposed method, there is no way to procure different capacity levels that, given clearing prices, achieves the same reliability at lower cost. That is, the ISO's approach allocates capacity purchases among the zones (including the ROS zone) cost-effectively.

■ **How ISO's Design Solves the Cost-Effectiveness Problem.** At a conceptual level, one might think that finding the cost-effective quantities at all possible clearing prices is a tedious process: For each set of possible prices (at least, whenever zones price separate), the possible quantities to demand in each zone need to be checked for cost-effectiveness, then modified until the right ratios are determined. Fortunately, there is instead a simple and general formula for determining the cost-effective quantities.

[12] For a graphical version of the ISO's proposed system and NNE demand curves using FCA10 inputs, see the ISO's December 10, 2015 Markets Committee presentation at http://www.iso-ne.com/static-assets/documents/2015/12/a09_iso_presentation_12_10_15.pptx. At the capacity levels in Example 3a, the system MRI = –.498 MWh/yr and the NNE MRI of capacity substitution is +.164 MWh/yr, so the total MRI of capacity additions in the NNE zone is –.498 + .164 = –.334 MWh/yr, as shown in the table. For the underlying demand curve and MRI data, see *op cit.*

The ISO's integrated set of zonal and system-level demand curves is built on a key insight: In order for capacity to be procured cost-effectively, the *relative* prices in any two capacity zones must equal their relative marginal reliability impacts. This is the same observation illustrated by the two ratios in the last Example 3a, above. In terms of formulas, this condition can be expressed as:

$$\frac{P_Z^{total}(Q_Z)}{P_{SYS}(Q_{SYS})} = \frac{MRIA_Z(Q_Z)}{MRI_{SYS}(Q_{SYS})} \qquad (10)$$

In this formula, the zonal price on the top-left is the zone's clearing price (*i.e.*, the sum of the system price and the zonal congestion price). The zonal MRI on the top-right is for capacity additions (*i.e.*, MRIA is the sum of the system MRI and the zonal MRI of capacity substitution). This formula is simply another way to express the ratios that we compared in Examples 1, 2, and 3 above: If the price ratio diverges from the MRI ratios, as in Examples 1 and 2, then the quantities demanded (at those prices) are not cost-effective – as we discovered directly. In Example 3, the price ratio equals the MRI ratio, and the quantities demanded are cost-effective.

Generalizing this observation, to be cost-effective a set of demand curves must be selected to satisfy equation (10). This might seem like a complex thing to do, but it is actually quite simple: It means the demand curves must be equal to the applicable MRI, times a constant. That's it. Why, you ask? In that way, the constants will cancel when the demand curves are used to determine prices (on the left-hand side of the price ratio above), and the cost-effectiveness condition – *by construction* – always holds. Stated more explicitly, the ISO's proposed demand curves for the system and the zones are cost-effective precisely because they are defined proportionately to the appropriate MRI:

$$P_{SYS}(Q_{SYS}) = Constant \times MRI_{SYS}(Q_{SYS})$$

And, for each zone,

$$Zonal\ Congestion\ Price(Q_Z) = Constant \times Zonal\ MRI(Q_Z)$$

where the *Zonal MRI* is for capacity substitution, so that $MRIA_Z(Q_Z) = Zonal\ MRI(Q_Z) + MRI_{SYS}(Q_{SYS})$. That is, plugging these formulas for the clearing prices into the left-hand side of equation (10) will result in the demand curves' constant canceling out in the ratios. That means the cost-effectiveness property holds at *all* prices, using the ISO's proposed demand curve formulas.

■ **Practical Implications.** Three key points are worth noting here. First, this analysis helps explain the ISO's concern with retaining the existing linear system demand curve. Achieving cost-effectiveness requires an integrated set of demand curves for each zone and the system (as the latter determines the ROS zone procurement). The existing linear demand curve does not satisfy the essential conditions that yield cost-effective procurement, and cannot be 'made to fit' these conditions.

At one level, this should not be surprising. The existing system demand curve was developed in isolation of any zonal demand curve considerations, at a time before the marginal reliability impact analyses the ISO has now conducted were available. Given the information now available, however, it is difficult to countenance continuing with a linear demand curve for which, at any prices, a different set of system and zonal demand curves would yield the same reliability at lower cost.

Second, cost-effectiveness matters in practice. For instance, with cost-effective demand curves, the FCA sends the right price signals for investment: They signal that investors should be willing to incur (say) 20% higher costs to build capacity in the ICZ (relative to the ROS) if, and only if, doing so will reduce total EENS by (at least) 20% more than building in the ROS. That is, the price signals will attract capacity in the more expensive zone when adding capacity there has a commensurately higher reliability benefit, and will not if it does not provide benefit commensurate to its higher cost.

Third, the cost-effectiveness condition explained in this section holds for the ISO's proposed demand curves at *all* prices, not just at Net CONE. For that reason, cost-effectiveness is a highly proscriptive condition on demand curve design: Cost-effectiveness admits only one class of demand curves, those which satisfy the ratio in (10). The only 'degree of freedom' in demand curve design, when cost effectiveness is a principle of the demand curve design, is the choice of the constant term in the demand curves.

This brings us to the remaining step of the ISO's demand curve methodology, which is the choice of the constant term in each demand curve. This constant corresponds to the penalty factor discussed earlier in this note, and its value is determined using the two other core design principles: Reliability and Sustainability (Principles 1 and 2). We explain the details of this final step next.

Applying the Reliability and Sustainability Principles

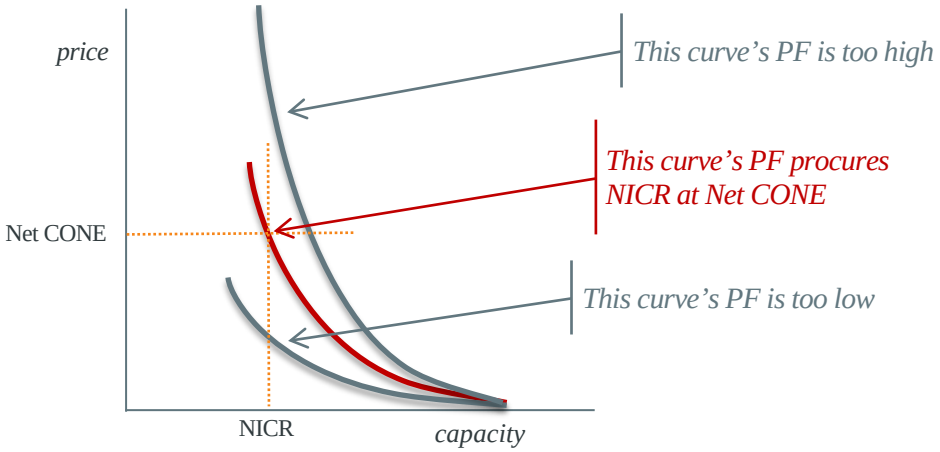
Although the general shape of the proposed capacity demand curves is determined by the marginal reliability impact functions (for the system and each zone, respectively), they also depend, in part, on the constant (the penalty factor) in each demand curve. Here, we explain how the ISO proposes to derive this parameter.

■ **Graphical Interpretation as a Scaling Factor.** The penalty factor described earlier in this memorandum can be thought of as a demand 'scaling' factor. Graphically, it simply 'stretches' the demand curves vertically. That is, if this scaling factor is increased by 10%, then the height (the prices) of the demand curve will increase by 10% at each quantity level; if the scaling factor doubles, the height of the demand curve at each quantity will double; and so forth.

The impact of three different demand scaling (penalty) factor values on a hypothetical system demand curve are illustrated on the next page (this figure omits the price cap, which is unaffected by the penalty factor). In an unconstrained system without capacity zones, the determination of the appropriate penalty factor value has a simple graphical interpretation: The penalty factor *PF* is set where the demand curve yields Net Cone at a capacity equal to NICR.

If the demand scaling factor is set too high, then the demand curves will procure more capacity than the reliability standards require, or set clearing prices above Net Cone, or both. At the opposite extreme, if the demand scaling factor is set too low, it will fail to remunerate investment at a level sufficient to meet the ISO's reliability planning obligations.[13]

[13] In determining the scaling (penalty) factor, the ISO uses the LOLE ≤ .1 reliability standard as proscribed in its existing reliability planning standards; it is not using a specific EENS value *per se* as the reliability requirement.



■ **Competitive Equilibrium and the Penalty (Demand Scaling) Factor.** Importantly, the penalty factor *PF* is not an *assumed* value. Rather, it is part of the demand curves and derived in a manner to satisfy the central design principles of Reliability and Market Sustainability. Specifically, the ISO proposes to set this demand scaling factor at the (smallest) value such that the demand curves for the system, and for each import-constrained zone, will satisfy these principles simultaneously.

These two principles are assessed by whether, at a price of Net Cone (in each zone), the demand curves yield sufficient capacity to ensure the annual Loss of Load Expectation (LOLE) is less than 0.1 for the system (respecting all zonal interface limits). In setting the penalty factor in this manner, we are making explicit use of the capacity market's long-run equilibrium property that the marginal capacity supply offer, whether existing or new, is offered at Net Cone. This long-run equilibrium is a property of the ISO's two-settlement capacity market design (also known as Pay for Performance), once the new design is completely phased in with its full Performance Payment Rate (starting with FCA15). This equilibrium bidding model is more consistent with the economics of competitive capacity supply pricing under the two-settlement capacity market than the historical bidding behavior of capacity suppliers, which was generally governed by their so-called 'going forward' (or avoidable) capital costs.

In addition, the assumption of Net Cone as the equilibrium capacity price is also consistent with the cash-flow and pricing assumptions actually employed in the ISO's discounted cash-flow models used to estimate Net Cone. In this way, the proposed method to set the demand curves' penalty (scaling) factor is internally consistent with both the current two-settlement capacity market design, as well as the Net Cone models with which the demand curves are calibrated.

■ **General Case with Equilibrium Price Separation.** In the general case, determining the scaling factor may involve an additional step beyond simply finding the point where the modified system demand curve intersects NICR and Net Cone. The additional step arises if there are different estimated Net Cone values in different capacity zones. In this case, the (single) scaling factor must be determined so that each zone's curve will clear, on average over time, at (at least) the zone's Net Cone value.

In practice, this involves two straightforward steps. First, we use the ISO's reliability planning model to compute the marginal reliability impact functions for each import constrained zone and for the system. Second, we determine the *smallest* value of the penalty factor such that, with the resulting system of demand curves, the system's overall LOLE ≤ .1, when evaluated accounting for all zonal import limits, at the quantities cleared under each zonal and system demand curve at Net Cone in each zone.

This means, in practice, that if there is a higher estimated ("administrative") Net Cone value for an import zone than the estimated Net Cone value applicable to the ROS zone, the scaling factor may be larger than would be obtained for an unconstrained system without capacity zones. In either case, however, the logic remains the same: The scaling factor is set just high enough – but no higher – than necessary to meet the 1-in-10 LOLE reliability criterion and remunerate capacity investment at Net Cone.