Many regions of the United States have deregulated electricity markets, where generation and distribution are handled by separate entities. Power plants generate electricity and sell it to **Load-Serving Entities** (LSEs), which then supply that electricity to end-use customers. The transactions between generators and LSEs typically occur through day-ahead and real-time auctions run by an **Independent System Operator** (ISO) or a **Regional Transmission Organization** (RTO). In this market, the relevant commodity is energy—that is, the actual physical electricity delivered to the grid.

To help ensure sufficient supply, several ISOs and RTOs also operate **capacity markets**. In a capacity market, the commodity traded is not energy itself, but rather the *obligation to sell energy in the future if needed*. When a generator's bid clears a capacity auction, the generator receives a fixed payment in exchange for committing to be available during a future delivery period. Importantly, if called upon to produce electricity during that period, the generator's actual energy sales still occur through the day-ahead and real-time markets. Thus, the capacity market helps an ISO/RTO ensure that enough resources will be available to meet projected future demand.

The rest of this document explores capacity markets in more depth. Subsequent sections explain the rationale for capacity markets, the structure of capacity auctions, the resource accreditation process, and alternatives to capacity markets.

I. Why Do Capacity Markets Exist?

A central goal of ISOs and RTOs is to ensure that sufficient resources are available to meet expected electricity demand. In New England, for example, regulators generally require that the system maintain enough generation and demand-side capacity to limit the probability of a **loss-of-load event**—a situation in which demand exceeds available supply—to no more than one event in ten years.

In the energy market, generators submit bids that specify both the quantity of energy they are willing to produce and the price at which they are willing to sell it. Grid operators then order these bids by price and draw on the cheapest units first. As a result, resources that are expensive to run and require high compensation—natural gas peaker plants, for example—will not be dispatched unless demand is unusually high. Because peak demand occurs only a few times per year, some expensive generators operate for very few hours annually, and other even-more-expensive generators may not run at all. Nevertheless, the system must ensure that these resources are available to meet demand during scarcity conditions and to cushion against unexpected demand spikes or supply disruptions.

In theory, energy-only markets—where generators are paid solely for the electricity they produce—should provide sufficient incentives for these resources to remain online. This is because even though a given generator may run for only a few hours in the year, the price that it receives in those hours should theoretically be high enough to reflect the fixed costs of being online for the rest of the year. Indeed, several grids operate according to this principle. ERCOT in Texas, for example, used

to let prices during scarcity conditions reach as high as $9,000 per MWh, which incentivized both new investment in generation (to capture these higher prices) and demand-side response (to avoid paying these higher prices). This system is generally agreed to result in reasonable reliability through strong price signals.

In practice, however, various constraints often prevent such pure price signalling. Many regions (including New England) have instituted energy price caps due to political pressure and the desire to protect consumers. Because price caps limit scarcity prices, expensive-to-run generators therefore cannot earn enough during rare high-demand periods to cover their fixed costs. This revenue shortfall, known as the "**missing money problem**," leads to underinvestment in generation and jeopardizes resource adequacy.

To solve the missing money problem, ISOs and RTOs such as ISO-NE have instituted **Forward Capacity Markets** (FCMs), which provide generators with fixed payments in exchange for a commitment to be available during future reliability events. These payments, determined by an auction held several years in advance, are supposed to make up the difference between the revenue required to stay online—i.e., the revenue that would be received absent price caps—and the actual revenue received given price caps. Thus, the FCM attempts to ensure that total compensation approximates what generators would earn in an unconstrained energy-only market.

FCMs also serve a planning function by signaling the need for new capacity. If expected supply falls below projected demand, market forces ensure that the fixed payments (clearing prices in the auction) grow larger, incentivizing investment in new generation. The capacity auction in New England has historically occurred three years before the "commitment period"—the time when generators receive payments and are expected to be dispatchable— to allow successful bidders to secure financing (eased by guaranteed payments) and construct new facilities before their obligation begins.

The following section explains how ISO-NE translates these abstract principles into practice.

II. How Do Capacity Markets Work?

The first stage of a FCM is the **Forward Capacity Auction** (FCA). In ISO-NE, FCAs are held annually, three years before the start of the capacity commitment period. Prior to the auction, ISO-NE estimates the total amount of capacity required to satisfy resource adequacy criteria. This estimate, which is effectively the market's demand curve for capacity, was historically fixed, i.e., a vertical line in a supply and demand graph. In the 2010s, however, ISO-NE adopted a *sloped demand curve* that allows the auction to clear within a small range of capacity, in order to reduce year-to-year price volatility.

Once the FCA begins, qualified participants (called "qualified capacity resources," or QCRs) submit bids that specify the price at which they are willing to provide capacity. These qualified participants include not only generators but also demand-response entities that can reduce net load

during peak conditions. ISO-NE then orders the bids by price, clearing the least-expensive bid until the capacity requirement is met. All cleared resources receive the price of the highest accepted bid under a uniform-price auction structure designed to reward low-cost resources.

Every participant that submits a clearing bid is awarded a **Capacity Supply Obligation** (CSO), which is a commitment to be available to supply electricity during the commitment period. Between the auction and the start of that period, ISO-NE monitors each resource's "critical path schedule" milestones (a development/upgrade roadmap) and may require demonstration of capability for new or upgraded resources. If a resource fails to meet readiness requirements, it must either shed its CSO or submit a restoration plan to regain compliance. During the commitment period, ISO-NE incentivizes compliance with CSOs during scarcity events through the **Pay For Performance** (PFP) system, which imposes penalties on generators that underperform and grants bonuses to those that exceed their obligations.

Finally, ISO-NE conducts a series of annual and monthly **reconfiguration auctions** between the initial FCA and the commitment period. These auctions allow the system to adjust total committed capacity in response to updated demand forecasts, and also enable participants to buy or sell CSOs to reflect changes in resource capability.

III. What Is Resource Accreditation?

To maintain reliability and determine how much capacity to procure through the FCM, ISO-NE must accurately model the grid's ability to meet peak demand. Simply counting the total installed megawatts of generation, however, is not sufficient. Different types of resources contribute differently to reliability depending on their operating characteristics, fuel sources, and likelihood of being available when demand is highest. A gas turbine that can start up quickly at any time, for example, matters more for reliability than a solar farm that produces little energy on a winter evening, even though both might have the same nameplate capacity. As a result, a system that has ample nameplate capacity could still struggle to meet a demand spike if many of its resources are unavailable at the same time. ISO-NE therefore needs a framework to estimate each resource's reliable contribution to meeting demand that factors in a given resource's unique characteristics. This framework is called **resource accreditation**.

Resource accreditation determines how much **qualified capacity** each resource may offer into the FCA. Qualified capacity represents the portion of a resource's nameplate capacity that ISO-NE expects to be available during demand peaks, essentially, measuring reliable rather than installed capacity. When ISO-NE considers whether it has procured enough supply to meet demand, it looks at qualified capacity and not nameplate capacity.

The qualified capacity of conventional thermal generators (gas, coal, oil, nuclear) is based on their **claimed maximum output** during peak demand conditions, known as **Seasonal Claimed**

Capability (SCC). These resources can generally operate whenever needed, so they can usually be relied upon to generate electricity during scarcity conditions. As such, qualified capacity for thermal generators is typically very close to nameplate capacity, which results in relatively high capacity revenues and thus a greater incentive for investment. Importantly, however, this value does *not* currently reflect thermal generators' forced-outage rates and fuel supply limitations.

Intermittent resources (wind, solar) are accredited very differently. Their qualified capacity is based on historical median performance during pre-defined **reliability hours**—periods when demand is high and the system is most stressed. This formula reflects the intermittency of these resources; it is irrelevant that a solar farm can theoretically supply 100MW if it can only supply 30MW during demand peaks. Qualified capacity for wind and solar, then, is typically a fraction of nameplate capacity. The current resource accreditation framework thus strongly favors dispatchable resources like natural gas.

Another, more sophisticated approach to resource accreditation—the one that ISO-NE is proposing to enact through its Capacity Auction Reforms (CAR) initiative—is marginal **Effective Load Carrying Capacity** (ELCC). To find a given resource's qualified capacity (called **Marginal Reliability Impact** (MRI)) with ELCC, ISO-NE would run a probabilistic model simulating future outages given expected weather, resource supply (including the specific resource in question), and energy demand, tweaking the supply and demand parameters until the desired reliability (one-in-ten loss of load) is reached. Then, ISO-NE would remove the specific resource and add perfectly-available capacity until the desired reliability is again reached. The amount of perfectly-available capacity added is the specific resource's qualified capacity. In simpler terms, ELCC measures the amount of additional load the system could serve with the resource (versus without it), while meeting the same target number of loss of load events. Ideally, this approach would also capture the fact that the reliability value of additional capacity declines as more of the same resource is added. For instance, once a large solar fleet is already meeting daytime demand, each new megawatt of solar provides progressively less reliability benefit, since those hours are no longer at risk.

IV. Alternatives to Capacity Markets

While capacity markets are currently the dominant mechanism for ensuring reliability in several U.S. regions, they are not the only approach. Other systems rely on energy-only markets or bilateral contracting to ensure that enough generation is available to meet demand.

An **energy-only market** (EOM) relies solely on revenues from energy and ancillary services to fund investment. There is no separate capacity payment; instead, high scarcity prices during tight supply intervals signal the need for new investment and demand-response, while also compensating for flexible resources. Texas' ERCOT system, for example, has historically allowed prices to rise as high as $9,000 per MWh, sending a strong signal that results in reasonable levels of reliability

(although ERCOT has since instituted a price cap). Proponents argue that this approach achieves reliability through pure price incentives, avoids the administrative complexity of capacity markets, and keeps consumers from giving extra revenue to resources that would have stayed online even absent capacity payments. Critics note, however, that high scarcity prices can be incredibly unpopular, leading to price caps that undermine the foundation of the system.

Other regions maintain reliability through **bilateral capacity contracts**. In these systems, load-serving entities are generally required to procure sufficient capacity through long-term contracts with generators. Alternatively, some regions where utilities are *vertically integrated*—i.e., where a single utility owns generation, transmission, and distribution infrastructure—require utilities to plan to build sufficient generation through **Integrated Resource Plans** (IRPs). The Western U.S. and much of the Southeast follow variants of this approach. Advantages include price stability, long-term investment certainty, and the ability to tailor procurement to local policy goals like renewables targets. The downside is that these systems limit competition, forcing regulators to make complex judgments about which resources to build or retain. These human decisions can be less efficient than market outcomes, especially given how understaffed some public utilities commissions are.