Wholesale Market Fundamentals: Energy and Capacity Markets Overview

In energy markets, generators are compensated for producing megawatt-hours. Generators submit offers reflecting the price at which they are willing to supply energy. The system operator stacks these offers from lowest to highest cost to meet forecasted demand for each hour. The highest-cost unit needed to meet demand sets the market-clearing price, and all dispatched units receive that price, reflecting the marginal unit.

This process reflects basic supply and demand. As demand rises, the operator must dispatch higher-cost units, which increases the clearing price. As demand falls, lower-cost units set the price. Scarcity pricing, occurring as prices rise (sometimes up to an administrative cap), signals the value of additional supply and in turn encourages new build.

Some regions supplement energy markets with capacity markets. Capacity markets pay resources for committing to be available during future peak periods. The product is not energy produced, but readiness to produce. Load-serving entities (e.g., utilities) must procure sufficient capacity to meet projected peak demand plus a reserve margin (effectively a buffer to account for forecast error). The price is set through a centralized auction, and this revenue stream helps ensure that resources which run infrequently but are critical during peak periods remain available during times of grid stress.

RTOs must maintain a balance across these resources, with supply equaling demand at every moment; any imbalances affect system frequency and can damage equipment or trigger outages. RTOs use economic dispatch to help maintain this balance, selecting the least-cost combination of available resources that can meet demand while respecting transmission limits and reliability constraints.

Rapid growth in large electricity loads can place upward pressure on retail electricity prices by tightening supply-demand balances. As higher-cost generation is dispatched to meet rising demand, wholesale prices increase, and those costs can flow through to all customers, not just the new load, much like prices rising in housing markets when demand outpaces supply. These effects are often transitional rather than permanent, with the magnitude and duration depending on how quickly supply can respond (through new generation, storage, transmission upgrades, or more flexible load) and whether regulatory, permitting, or interconnection barriers delay that response. Moreover, the effects depend on how costs are allocated, both to ensure new loads pay in line with the costs they impose on the system and to potentially create incentives for large loads to minimize their grid impacts, such as through time-of-use pricing or load flexibility, to reduce costs.