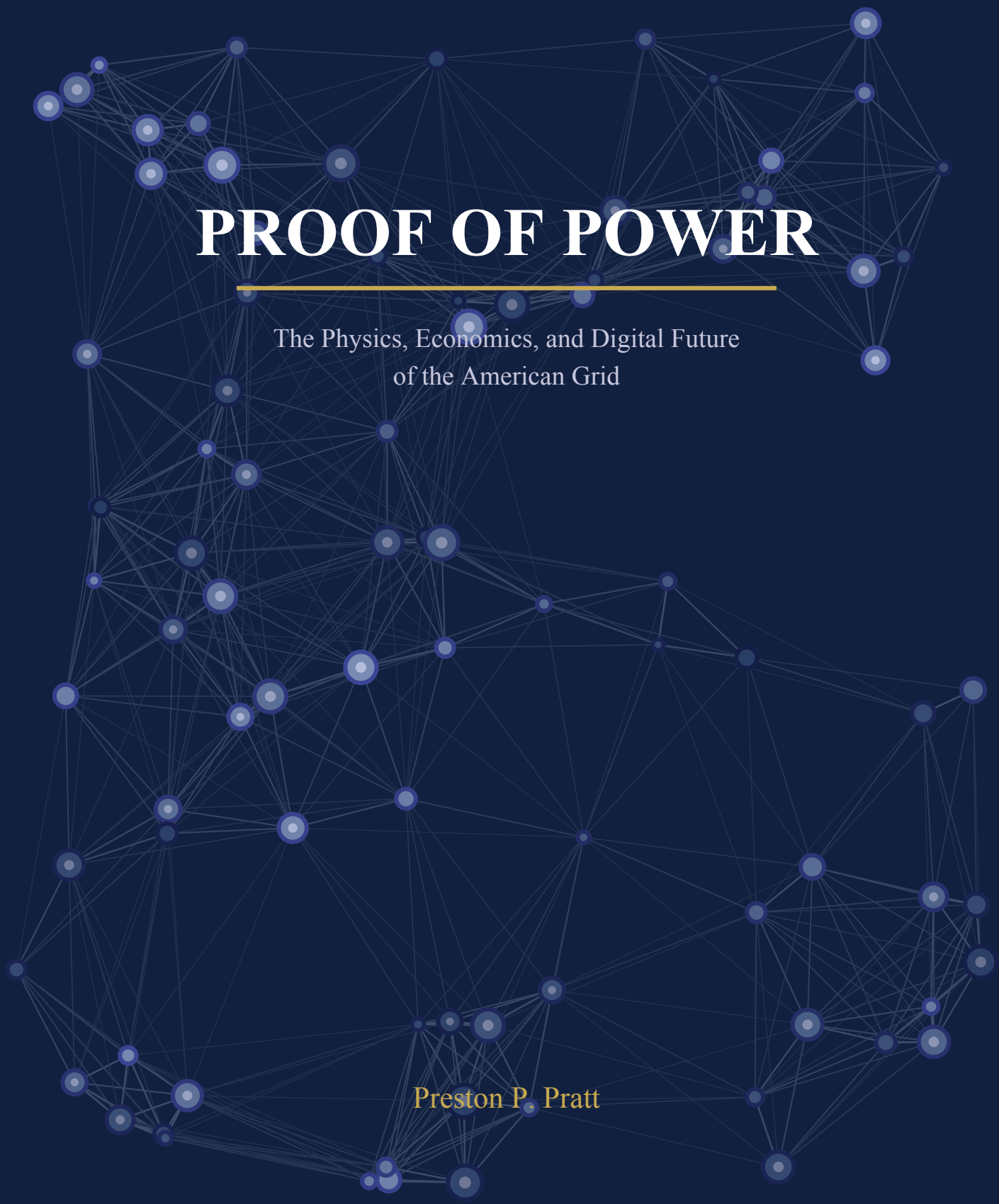

PROOF OF POWER

The Physics, Economics, and Digital Future of the American Grid

Preston P. Pratt

Connect with the Author

prestonppratt.com

Blog: blog.prestonppratt.com
LinkedIn: linkedin.com/in/prestonppratt
X / Twitter: x.com/prestonppratt
Facebook: facebook.com/prestonppratt

Published by Pratt Materials • prattmaterials.com

For bulk orders, speaking inquiries, and media requests, please visit prestonppratt.com.

Proof of Power

Proof of Power

The Physics, Economics, and Digital Future
of the American Grid

Preston P. Pratt

This book is intended for informational and educational purposes only.
It does not constitute engineering, security, legal, or investment advice.

First Edition

ISBN 000-0-00-000000-0 (hardcover)
ISBN 000-0-00-000000-0 (ebook)

Printed in the United States of America

For those who keep the lights on.

Contents

Preface

This book is an attempt to make legible, from the ground up, the most complex machine ever built by human hands: the American electric power grid. It is written for the reader who wants to understand not only what the grid is — its physical components, its institutional architecture, its market mechanisms — but *why* it is the way it is. Every market rule, every regulatory boundary, every pricing mechanism examined in these pages exists because someone, somewhere, was trying to solve a problem created by the unyielding physics of electricity, the legacy of a century of institutional development, or both.

The book is organized in six parts, each building upon the last.

Part I: The Physical Foundation (Chapters 1–2) begins where any serious study of the grid must begin — with physics. Chapter 1 introduces the three tiers of the electric power system: generation, transmission, and distribution. It explains the synchronous frequency that binds the grid into a single machine, the supply-demand balance that must be maintained every instant, and the technical characteristics of the generation technologies that supply the nation's electricity. Chapter 2 extends this physical foundation to the continental scale, examining the three great interconnections — Eastern, Western, and ERCOT — that divide the North American grid into distinct synchronized regions, and the balancing authorities that maintain real-time equilibrium within them.

Part II: Regulation and Ownership (Chapters 3–4) turns from physics to institutions. Chapter 3 examines the vertically integrated monopoly utility — the organizational form that dominated the American electricity industry for most of the twentieth century — and the regulatory compact of cost-of-service ratemaking that governed it. Chapter 4 surveys the public power sector: the Tennessee Valley Authority, the Bonneville Power Administration and other federal power marketing administrations, rural electric cooperatives, and municipal utilities — entities that provide electricity outside the investor-owned utility model and that collectively serve roughly a quarter of the nation's load.

Part III: Markets and Deregulation (Chapters 5–6) traces the transformation of wholesale electricity from a regulated bilateral commodity to a product traded in organized competitive markets. Chapter 5 tells the story of FERC Orders 888 and 2000, the creation of Regional Transmission Organizations and Independent System Operators, and the emergence of retail choice in restructured states. Chapter 6 examines the mechanics of these markets in detail — Locational Marginal Pricing, security-constrained economic dispatch, the design of capacity markets, and the persistent "missing money" problem.

Part IV: Regional Profiles and Comparative Analysis (Chapters 7–9) provides detailed

examinations of the major regional electricity systems. Chapter 7 takes on ERCOT and the Texas energy-only market experiment, including the catastrophic stress test of Winter Storm Uri. Chapter 8 examines the Western Interconnection, the California ISO, and the Energy Imbalance Market. Chapter 9 profiles the Northeast and Midwest corridors — ISO New England, NYISO, MISO, and SPP — exploring the contrasting archetypes of dense urban load pockets and remote rural generation.

Part V: Transformation and Resilience (Chapters 10–11) looks forward to the forces reshaping the grid from within. Chapter 10 examines the physics and policy of grid decarbonization: the transition from synchronous machines to inverter-based resources, the integration of distributed energy resources through FERC Order 2222, and the policy landscape shaped by the Inflation Reduction Act and state clean energy mandates. Chapter 11 addresses the twin threats to the grid's survival — cybersecurity and physical resilience — from NERC CIP standards and nation-state cyber adversaries to extreme weather, geomagnetic disturbances, and electromagnetic pulse.

Part VI: The Digital Frontier (Chapters 12–13) examines the new categories of load reshaping the grid from the demand side. Chapter 12 examines the explosive growth of data centers — driven by cloud computing and artificial intelligence — as a new category of massive, reliability-demanding load that is reshaping utility planning, transmission development, and generation investment after decades of essentially flat demand growth. Chapter 13 examines Bitcoin mining as a strikingly different kind of digital load — one whose unique physical properties of interruptibility, location-agnosticism, and price-elasticity allow it to serve not merely as a consumer of grid services but as a provider of them, through demand response, the monetization of curtailed renewable energy, and flare gas mitigation.

The book concludes with three appendices — a Comparative Reference Table of the major RTOs and ISOs, a Glossary of Key Terms, and a comprehensive Bibliography — followed by an Index, providing quick-reference resources for the concepts, institutions, and sources examined throughout the text.

A note on scope: this book focuses on the contiguous United States, with references to Canadian provinces and Mexico where their grids interconnect with American systems. It does not attempt to address the electricity systems of Hawaii, Alaska, or U.S. territories in detail, though the catastrophe of Hurricane Maria in Puerto Rico is examined in Chapter 11 as a case study in grid resilience.

The reader who works through these chapters in sequence will find that each builds upon the last — that the market structures of Part III are responses to the physics of Part I, that the regional profiles of Part IV are variations on the institutional themes of Parts II and III, and that the future challenges of Parts V and VI are the consequences of everything that came before. The grid is a system, and understanding it requires engaging with all of its dimensions — physical, economic, legal, and political — simultaneously. Chapters 12 and 13, on data centers and Bitcoin mining, illustrate this principle in microcosm: two categories of digital load that engage every dimension of grid architecture examined in this book — physics, market design, transmission planning, regulatory policy, and the economics of generation investment — and that together embody the central paradox of the modern grid, in which new demand can simultaneously stress the system and strengthen it.

* * *

Part I

The Physical Foundation

Chapter 1: The Physics of the Triple-Tier System

Electricity is unlike any other commodity traded in modern markets. It cannot be economically stored at scale, it travels at nearly the speed of light, and it must be produced at the precise instant it is consumed. These physical realities are not mere engineering curiosities — they are the foundational constraints upon which every market rule, every regulatory framework, and every policy decision in the American power sector is ultimately built. One cannot understand why wholesale electricity markets clear every five minutes, why transmission planning disputes consume decades of regulatory proceedings, or why the integration of renewable energy demands fundamental institutional adaptation without first understanding the physics that make the electric grid the largest and most complex machine ever constructed by human civilization.

This chapter introduces the three physical tiers of that machine: generation, transmission, and distribution. But before examining each tier in turn, we must begin with the phenomenon that binds them into a single, interdependent system — the synchronous frequency that pulses through the grid sixty times per second, every second of every day, connecting hundreds of thousands of miles of conductor and thousands of generators into one coordinated whole.

The reader who internalizes the physical principles in this chapter will find that the market structures, regulatory institutions, and policy debates examined in subsequent chapters are not arbitrary constructs but rather logical — indeed, often inevitable — responses to the unyielding demands of electrical physics.

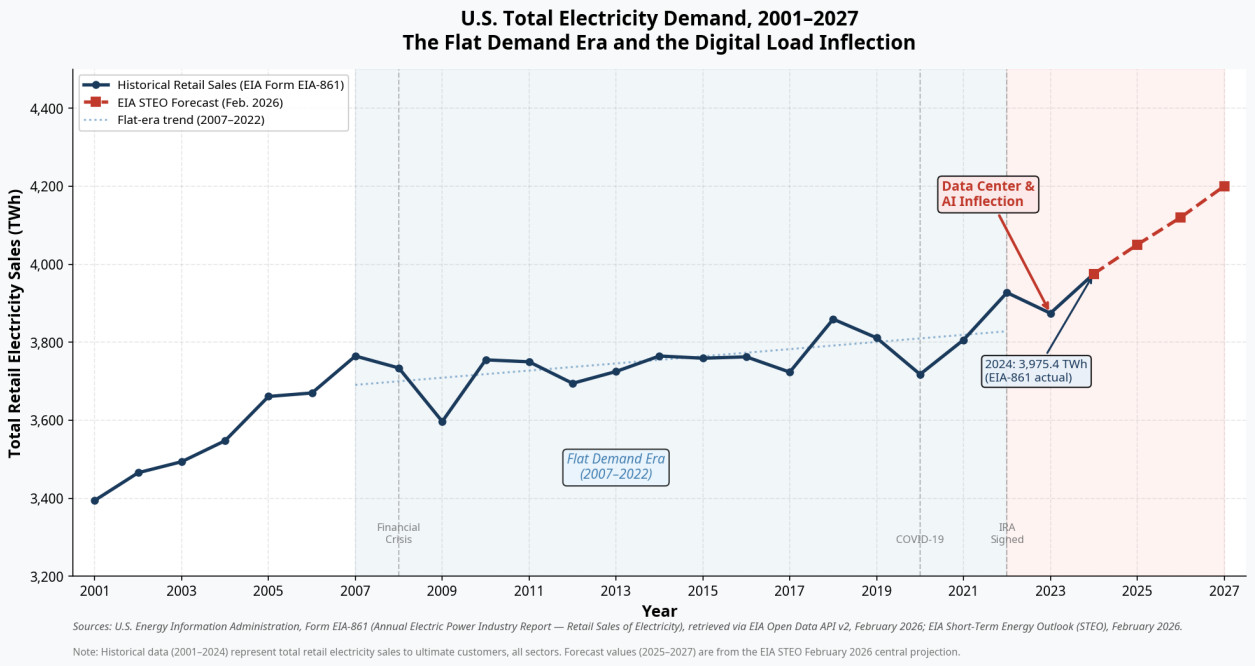

Figure 1.0: U.S. Total Electricity Demand, 2001–2027 — The Flat Demand Era and the Digital Load Inflection (Source: EIA Form EIA-861, EIA STEO Feb. 2026)

* * *

1.1 The Synchronous Machine: Frequency, Balance, and Inertia

1.1.1 The Sixty-Hertz Heartbeat

The alternating current that powers the American grid oscillates at a nominal frequency of 60 hertz — sixty complete sinusoidal cycles per second. This frequency is not an arbitrary convention. It was standardized in the early twentieth century as a practical compromise: high enough to eliminate perceptible flicker in incandescent lighting, low enough to permit efficient transformer design and limit reactive losses in long-distance transmission. Europe, following different engineering traditions, settled on 50 hertz. Once established, these standards became locked in by the enormous capital stock of generators, motors, transformers, and end-use devices designed around them.

But the 60-hertz frequency of the American grid is not merely a design specification stamped on equipment nameplates. It is a real-time physical measurement of the instantaneous balance between the mechanical energy being fed into the system by generators and the electrical energy being withdrawn by loads. Understanding this relationship is essential to grasping why the grid operates as it does.

1.1.2 The Physics of Synchronous Generation

The vast majority of electricity in the United States has historically been produced by synchronous generators — machines in which a magnetized rotor spins inside a stationary armature (the stator) wound with copper conductors. As the magnetic field of the rotor sweeps past the stator windings, it induces an alternating electromotive force in accordance with Faraday's law of electromagnetic induction. The frequency of the resulting alternating current is determined by a straightforward relationship:

$$f = (N \times P) / 120$$

where f is the electrical frequency in hertz, N is the rotational speed of the rotor in revolutions per minute, and P is the number of magnetic poles on the rotor. A two-pole generator must spin at 3,600 RPM to produce 60-hertz power; a four-pole machine achieves the same frequency at 1,800 RPM.

This equation reveals something profound: the electrical frequency of the grid is a direct, physical manifestation of the rotational speed of the generators connected to it. When we say the grid is operating at 60 hertz, we are equivalently saying that every synchronous generator on the system is spinning at precisely its synchronous speed. The frequency is not set by a clock or a control signal — it emerges from the aggregate mechanical rotation of thousands of coupled machines.

1.1.3 The Supply-Demand Balance

The coupling between generator speed and system frequency creates the grid's most fundamental operating constraint: supply must equal demand at every instant. The mechanism is elegantly simple in principle, though enormously complex in practice.

Consider a single synchronous generator connected to a load. The prime mover — a steam turbine, a gas turbine, a hydro turbine — applies mechanical torque to the rotor. The electrical load connected to the stator creates an opposing electromagnetic torque. When these two torques are exactly equal, the rotor spins at constant speed and the frequency is stable.

Now suppose demand increases — a factory starts its motors, a million air conditioners cycle on across a sweltering city. The additional electrical load increases the electromagnetic torque opposing the rotor's rotation. If the mechanical input torque from the prime mover does not change instantaneously, the rotor begins to decelerate. This deceleration is not a malfunction; it is the physics of Newton's second law applied to a rotating mass. The rotor slows, and because frequency is directly proportional to rotational speed, the system frequency drops.

Conversely, if a large load suddenly disconnects — a major industrial facility trips offline, or a transmission line carrying power to a load center opens — the mechanical torque driving the generators momentarily exceeds the electrical torque opposing them. The rotors accelerate, and the frequency rises.

This relationship can be expressed through the swing equation, the fundamental differential equation governing the dynamics of a synchronous machine:

$$2H \times (d\omega/dt) = P_{mechanical} - P_{electrical}$$

where H is the inertia constant of the machine (expressed in seconds), ω is the angular velocity, $P\_mechanical$ is the mechanical power input from the prime mover, and $P\_electrical$ is the electrical power output. This equation governs not only individual machines but, in aggregate form, the behavior of the entire interconnected system.

1.1.4 Frequency Deviation and Its Consequences

Under normal operating conditions, the frequency of the Eastern Interconnection — the vast synchronous network spanning from the Rocky Mountains to the Atlantic seaboard, from Saskatchewan to Florida — varies by only tiny fractions of a hertz around the 60-hertz nominal. A deviation of even 0.05 hertz reflects a significant imbalance between generation and load across the entire interconnection.

The consequences of frequency deviation escalate with magnitude. Small deviations are corrected automatically by governor response and automatic generation control, mechanisms we will examine in later chapters on ancillary services markets. Larger deviations trigger progressively more serious protective actions. At approximately 59.95 hertz, grid operators issue alerts. Below 59.5 hertz, utilities begin automatic under-frequency load shedding — the deliberate, involuntary disconnection of customer load in predetermined blocks to arrest the frequency decline. If frequency falls below approximately 57 hertz, generators begin tripping offline to protect their turbine blades from resonant vibration damage, potentially initiating a cascading failure. Below roughly 55 hertz, a complete system collapse — a blackout — becomes nearly inevitable.

The catastrophic blackout of August 14, 2003, which plunged fifty-five million people across the northeastern United States and Ontario into darkness, illustrated the speed and violence with which cascading failure can propagate through a synchronous system. The entire collapse, from the initial contingency to complete system separation, unfolded in approximately nine seconds.

On the high side, over-frequency events are less common but equally dangerous. Excess generation relative to load causes generators to accelerate, potentially damaging turbine blades, exceeding bearing tolerances, and stressing generator windings beyond their design limits.

1.1.5 Inertia: The Grid's Shock Absorber

The reason the grid does not collapse with every momentary imbalance between supply and demand is inertia — the kinetic energy stored in the massive rotating masses of synchronous generators and their prime movers. A large steam turbine generator set may have a rotor weighing several hundred tons spinning at 1,800 or 3,600 RPM. The kinetic energy stored in this rotating mass is enormous, and it acts as a buffer: when electrical demand suddenly exceeds supply, the deficit is initially met by converting the kinetic energy of the rotors into electrical energy. The rotors slow down — frequency drops — but they do so gradually, buying time for control systems to increase mechanical input power and restore balance.

The inertia constant H, expressed in units of megawatt-seconds per megavolt-ampere (or

equivalently, seconds), represents the time a generator could supply its rated output purely from its stored kinetic energy. Typical values range from two to ten seconds depending on machine type: large steam turbines at the higher end, smaller gas turbines and hydro units at the lower end. For the system as a whole, the aggregate inertia determines the rate of frequency change (the "rate of change of frequency," or ROCOF) following a disturbance. Higher system inertia means a slower rate of frequency change, affording more time for corrective action.

This concept has taken on critical importance in the twenty-first century because the two fastest-growing generation technologies — wind and solar photovoltaics — do not inherently provide synchronous inertia. Wind turbines connected through power electronic converters and solar panels connected through inverters are decoupled from the grid's rotational frequency. They can be controlled to mimic some inertial response (so-called "synthetic inertia" or "fast frequency response"), but they do not naturally contribute the kinetic energy buffer that synchronous machines provide. As these resources displace conventional generators, the aggregate inertia of the system declines, making the grid more susceptible to rapid frequency excursions. This emerging challenge — managing a low-inertia grid — is reshaping not only engineering practice but also market design and regulatory policy, as subsequent chapters will explore.

* * *

1.2 Generation: A Technical Taxonomy

1.2.1 The Hierarchy of Dispatch

Not all generators are created equal. They differ in their fuel costs, their ability to start up and shut down, their flexibility to change output levels, and their suitability for sustained versus intermittent operation. These physical characteristics give rise to a natural hierarchy that has organized generation planning and operations for decades: baseload, intermediate (or load-following), and peaking resources.

This framework, while increasingly complicated by the growth of zero-marginal-cost renewables, remains indispensable for understanding how the physical characteristics of generating technologies shape their economic roles.

1.2.2 Baseload Generation

Baseload resources are designed to operate continuously at or near full output for extended periods. They are characterized by high capital costs, low variable (fuel) costs, limited operational flexibility, and high capacity factors.

Nuclear power represents the archetypal baseload resource. A modern pressurized water reactor or boiling water reactor operates at a heat rate of approximately 10,400 BTU per kilowatt-hour — roughly 33 percent thermal efficiency — but its fuel costs are so low (on the order of $5 to $8 per megawatt-hour) that it is economical to run at maximum output virtually all the time. Nuclear plants are typically designed for capacity factors of 90 percent or higher. They are slow to start (requiring days to bring from cold shutdown to full power), difficult to cycle (thermal stresses on fuel cladding and reactor pressure vessels make frequent power changes undesirable), and extraordinarily expensive to build (capital costs of $6,000 to $12,000 per kilowatt for recent U.S. projects). Once running, however, they produce enormous quantities of carbon-free electricity at very low marginal cost.

Coal-fired generation, historically the backbone of American baseload supply, operates on the Rankine steam cycle at heat rates of roughly 8,500 to 10,500 BTU per kilowatt-hour for modern supercritical units (corresponding to thermal efficiencies of 32 to 40 percent). Coal plants were traditionally designed for continuous operation at high capacity factors, though the combination of low natural gas prices and environmental regulations has pushed many into cycling duty for which they were not designed, accelerating mechanical wear and increasing maintenance costs. The marginal cost of coal-fired generation — dominated by fuel costs — has historically ranged from $20 to $40 per megawatt-hour depending on coal prices, plant efficiency, and the cost of emissions allowances where applicable.

Large hydroelectric facilities with substantial reservoir storage can also serve as baseload resources, offering heat rates of zero (no thermal conversion), negligible marginal costs, and capacity factors ranging widely depending on hydrology and reservoir management constraints.

1.2.3 Intermediate and Load-Following Generation

Between the always-on baseload units and the quick-start peakers sits a category of generation designed to follow the daily rise and fall of demand. These intermediate resources must be able to increase and decrease output across a wide operating range, start and stop with reasonable speed, and operate economically at varying capacity factors.

Natural gas combined-cycle (NGCC) plants have come to dominate this role. A combined-cycle plant pairs one or more gas turbines (operating on the Brayton cycle) with a heat recovery steam generator and a steam turbine (operating on the Rankine cycle), capturing waste heat from the gas turbine exhaust to produce additional electricity. This configuration achieves heat rates of 6,300 to 7,500 BTU per kilowatt-hour — thermal efficiencies of 45 to 54 percent for the most advanced units — making it the most efficient fossil fuel technology in widespread commercial use. Modern NGCC plants can ramp at rates of 15 to 30 megawatts per minute, start from a hot condition in one to two hours, and cycle daily without excessive mechanical penalty. At natural gas prices of $2 to $4 per million BTU, the marginal cost of NGCC generation falls in the range of $15 to $30 per megawatt-hour, which has made these plants the marginal price-setting units in many American wholesale markets.

The NGCC plant's combination of relatively low capital cost ($800 to $1,300 per kilowatt),

moderate heat rates, operational flexibility, and lower carbon intensity compared to coal has made it the generation technology of choice for the past two decades. It is the Swiss army knife of the modern American power fleet.

1.2.4 Peaking Generation

Peaking resources exist to serve demand during the highest-load hours and to provide reserves against contingencies. They are characterized by low capital costs, high variable costs, rapid start capability, and low capacity factors (often below 10 to 15 percent annually).

Natural gas combustion turbines (simple-cycle gas turbines) are the quintessential peaker. Operating on the Brayton cycle alone — without the waste heat recovery of a combined-cycle configuration — they achieve heat rates of 9,000 to 11,500 BTU per kilowatt-hour (30 to 38 percent efficiency). They are expensive to run but cheap to build ($400 to $700 per kilowatt) and can start from cold conditions in ten to twenty minutes, reaching full output within thirty minutes. Some modern aeroderivative gas turbines — adapted from jet engine designs — can start and reach full load in under ten minutes with ramp rates exceeding 30 megawatts per minute.

The economic logic of peaking resources illustrates a principle that recurs throughout electricity markets: because demand is highly variable but supply must always meet it, the system requires resources that may run very few hours per year but must be available when called upon. The cost recovery challenge this creates — compensating generators that provide capacity but generate relatively little energy — is one of the central design problems of wholesale electricity markets, examined in detail in Part II of this book.

1.2.5 Pumped-Storage Hydroelectricity: The Original Grid Battery

Long before lithium-ion batteries became a grid resource, the electric utility industry developed a technology for storing energy at enormous scale: pumped-storage hydroelectricity. A pumped-storage facility consists of two reservoirs at different elevations connected by tunnels, turbines, and pumps. During periods of low electricity demand and low prices — typically overnight — the facility pumps water from the lower reservoir to the upper reservoir, consuming electricity. During periods of high demand and high prices, the water is released back through turbines to generate electricity, much like a conventional hydroelectric dam. The facility is, in essence, a rechargeable battery built from water and gravity.

The scale of these facilities is difficult to overstate. The United States has approximately 22 gigawatts of pumped-storage capacity spread across roughly 40 facilities — making pumped hydro, by a wide margin, the largest form of energy storage on the American grid. By comparison, all utility-scale battery storage in the country totaled approximately 21 gigawatts by 2024, a figure that has only recently approached the pumped-hydro fleet that was largely built between the 1960s and 1980s.

Several pumped-storage facilities deserve particular attention for their engineering ambition and

their operational importance. The Bath County Pumped Storage Station in Virginia, operated by Dominion Energy, is the largest pumped-storage facility in the world. Its six reversible Francis turbines can generate 3,003 megawatts — roughly the output of three nuclear reactors — and its upper reservoir, nestled in the Allegheny Mountains at an elevation of 2,846 feet, holds approximately 35 billion gallons of water. Bath County can run at full output for approximately eleven hours before its upper reservoir is depleted, providing critical peaking capacity and grid stabilization services to the PJM Interconnection.

The Tennessee Valley Authority's Raccoon Mountain Pumped-Storage Plant, located atop Raccoon Mountain near Chattanooga, Tennessee, exemplifies the elegant engineering of these facilities. Built between 1970 and 1978, Raccoon Mountain's upper reservoir sits 1,000 feet above the Tennessee River, connected by a 2,100-foot tunnel drilled through solid rock. Its four generating units can produce 1,616 megawatts, enough to power roughly 1.2 million homes during peak demand. The facility can transition from full pumping to full generation in minutes — a flexibility that makes it invaluable for balancing the TVA system, which increasingly relies on variable wind and solar generation.

The Ludington Pumped Storage Power Plant in Michigan, operated by Consumers Energy, occupies a 1,000-acre reservoir perched on bluffs overlooking Lake Michigan. Its six reversible turbines generate 1,872 megawatts, making it one of the largest in the world. Ludington uses Lake Michigan itself as its lower reservoir — a virtually limitless water supply that eliminates the evaporation and capacity constraints that affect closed-loop systems. The facility provides critical peaking power and frequency regulation to the Midcontinent Independent System Operator (MISO).

Pumped-storage facilities share a crucial characteristic with the synchronous generators described in Section 1.1: their turbines are rotating machines that contribute inertia to the grid. When operating in generation mode, a pumped-storage unit behaves like any other synchronous generator, its massive rotor spinning in lockstep with the grid frequency and contributing rotational kinetic energy that resists sudden frequency deviations. This makes pumped storage doubly valuable in a grid transitioning toward inverter-based resources — it provides not only energy storage and peaking capacity but also the physical grid stability services that wind and solar cannot.

The round-trip efficiency of pumped-storage facilities — the ratio of energy generated to energy consumed — typically ranges from 70 to 85 percent. This means that for every 100 megawatt-hours of electricity used to pump water uphill, approximately 70 to 85 megawatt-hours are recovered when the water flows back down. While this represents an energy loss, the economic value is positive because the electricity consumed during off-peak hours is far cheaper than the electricity generated during peak hours. The price spread between off-peak and on-peak electricity — a spread that has widened as solar generation depresses midday prices and evening demand ramps create steep price increases — determines the profitability of pumped-storage operations.

Despite the proven effectiveness of pumped storage, very few new facilities have been built in the United States since the 1990s. The primary barriers are environmental permitting (constructing large reservoirs on mountainsides faces significant opposition), long construction timelines (typically seven to ten years), high upfront capital costs, and the availability of suitable topography. However, the growing need for long-duration storage — storage that can discharge for hours or days rather than the two to four hours typical of lithium-ion batteries — has renewed interest in pumped hydro. Several new projects are

in various stages of permitting and development, and the Department of Energy has identified pumped storage as a critical component of a decarbonized grid.

1.2.6 Variable Renewable Resources

Wind and **solar photovoltaic (PV)** generation have grown from negligible shares to providing approximately 15 percent of total U.S. electricity as of the mid-2020s, a fraction that continues to grow rapidly. These resources defy the traditional baseload/intermediate/peaking taxonomy because their output is governed by meteorological conditions rather than operator dispatch decisions.

Wind turbines convert kinetic energy from moving air into electricity through aerodynamic lift on turbine blades coupled to a generator (increasingly through power electronic converters rather than direct synchronous coupling). Capacity factors for modern onshore wind installations range from 25 to 45 percent depending on site quality, with the best Great Plains locations exceeding 50 percent. Offshore wind projects achieve capacity factors of 40 to 55 percent owing to stronger and more consistent marine winds.

Solar PV converts photons directly into electricity through the photovoltaic effect in semiconductor materials (predominantly crystalline silicon). Capacity factors for utility-scale solar range from 20 to 30 percent, with the highest values in the desert Southwest. Solar output follows a predictable diurnal pattern — zero at night, peaking around solar noon — but is subject to short-term variability from cloud passage.

Both technologies have zero marginal fuel cost and are dispatched whenever available, placing them at the bottom of the merit order. This characteristic has profound implications for wholesale market prices, for the utilization of conventional generators, and for system operations — implications that form a recurring theme throughout this book.

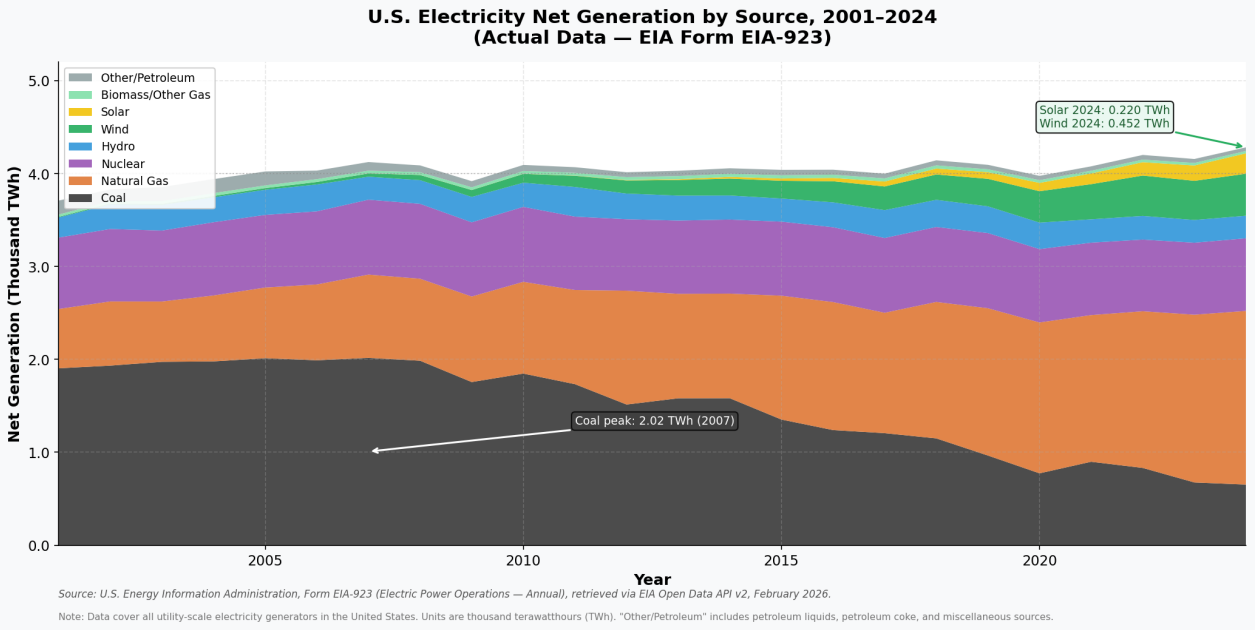

Figure 1.1: U.S. Electricity Net Generation by Source, 2001–2024 (Source: EIA Form EIA-923)

1.2.7 Merit Order Dispatch

The economic dispatch of generating resources follows the merit order: in any given interval, the system operator (or the market) ranks available generators from lowest to highest marginal cost and dispatches them in that order until supply equals demand. Resources with zero or near-zero marginal cost — nuclear, wind, solar, run-of-river hydro — are dispatched first. NGCC plants typically fall in the middle of the stack. Simple-cycle gas turbines and oil-fired units, with the highest marginal costs, are dispatched last, only when demand is high enough to require them.

The marginal cost of the most expensive unit dispatched in any interval sets the system marginal price in energy markets — the price paid to all dispatched generators. This "uniform clearing price" mechanism is the foundation of organized wholesale electricity markets, and its interaction with zero-marginal-cost renewables is reshaping price formation in ways that challenge long-standing market design assumptions.

* * *

1.3 The Transmission "Highway": High-Voltage Power Delivery

1.3.1 Why High Voltage?

The fundamental reason for high-voltage transmission is contained in a single equation from elementary circuit theory. The power dissipated as heat in a conductor — the resistive loss, colloquially known as "line loss" — is:

$$P_{loss} = I^2 \times R$$

where I is the current flowing through the conductor and R is the resistance. Since power transmitted equals voltage times current ($P = V \times I$), for a given amount of power, increasing the voltage proportionally decreases the current. Because losses scale with the *square* of the current, doubling the voltage cuts losses by a factor of four.

This relationship explains the entire architecture of the transmission system. A large generating station may produce power at 13.8 to 25 kilovolts (kV). A step-up transformer at the plant raises this to transmission voltage — 138 kV, 230 kV, 345 kV, 500 kV, or 765 kV — reducing the current by the same factor and making it feasible to move large quantities of power over hundreds of miles with tolerable losses. Without high-voltage transmission, the resistive losses would be so enormous that electricity could only be consumed within a few miles of where it was generated — indeed, this was precisely the limitation of Thomas Edison's original direct-current distribution system in the 1880s.

1.3.2 The Transformer: Enabling Technology

The device that makes high-voltage AC transmission possible is the transformer, whose operating principle is again rooted in Faraday's law. A transformer consists of two coils of wire (the primary and secondary windings) wound around a shared magnetic core, typically made of laminated silicon steel. An alternating current in the primary winding creates a time-varying magnetic flux in the core, which induces an alternating voltage in the secondary winding. The ratio of secondary to primary voltage equals the ratio of turns in the secondary winding to turns in the primary winding:

$$V_{secondary} / V_{primary} = N_{secondary} / N_{primary}$$

Modern power transformers achieve efficiencies of 99 percent or higher, making voltage transformation nearly lossless. A large substation transformer rated at 500 MVA (megavolt-amperes) may stand two stories tall, weigh 400 tons, contain 10,000 gallons of mineral oil for cooling and insulation, and cost $5 million to $15 million. These transformers are custom-built with lead times of twelve to eighteen months, and their physical size and weight make transportation — often requiring specialized rail cars or heavy-haul trucks — a significant logistical undertaking. The long replacement lead times for large power transformers represent a significant grid resilience vulnerability, a topic of growing concern in homeland security and grid reliability discussions.

It is worth noting that transformers operate on the principle of electromagnetic induction, which requires a *changing* magnetic flux — a condition naturally satisfied by alternating current. Direct current, being constant, does not induce voltage in a transformer. This fundamental physical fact is why Nikola

Tesla and George Westinghouse's alternating current system triumphed over Edison's direct current system in the "War of Currents" of the 1890s, and why the overwhelming majority of the world's power systems operate on AC to this day.

1.3.3 Transmission Voltage Levels and the Network Hierarchy

The American transmission system operates at a hierarchy of voltage levels, each serving a distinct function:

Extra-high voltage (EHV): 345 kV, 500 kV, 765 kV. These lines form the backbone of the bulk power system, moving large quantities of power over long distances between regions. A single 765-kV line can carry 2,000 to 2,400 megawatts — enough to serve a city of well over a million people. The 765-kV system, developed by American Electric Power in the 1960s, is concentrated in the Ohio Valley and mid-Atlantic regions. The 500-kV system is widespread across the Southeast, West, and mid-Atlantic. Lines at 345 kV are common throughout the Eastern Interconnection.

High voltage (HV): 69 kV, 115 kV, 138 kV, 230 kV. These lines serve as the arterial network, connecting load centers to the EHV backbone, delivering power to large industrial customers, and interconnecting generation plants with the broader network. The 230-kV level is the workhorse of many regional systems, particularly in the West.

The distinction between "transmission" and "distribution" is not merely technical but carries profound regulatory significance. Under the Federal Power Act, the Federal Energy Regulatory Commission (FERC) exercises jurisdiction over transmission in interstate commerce, while distribution facilities fall under state regulatory authority. The physical boundary between these jurisdictions — typically defined by voltage level, function, or both — has been the subject of extensive regulatory proceedings and litigation, as we will examine in Part III.

1.3.4 AC Versus HVDC Transmission

While the vast majority of the American transmission system operates on alternating current, high-voltage direct current (HVDC) transmission plays a specialized but increasingly important role.

AC transmission benefits from the ease of voltage transformation and the natural zero-crossing of the alternating waveform, which facilitates fault current interruption by circuit breakers. However, AC transmission suffers from several limitations on long lines: reactive power losses (energy oscillating between electric and magnetic fields rather than being delivered to loads), stability limits (the maximum power transferable decreases with increasing line length as the angular separation between sending and receiving ends grows), and skin effect (the tendency of alternating current to flow on the surface of conductors, increasing effective resistance).

HVDC transmission eliminates these AC-specific problems. Direct current has no reactive power, no stability limit related to angular separation, and no skin effect. Furthermore, HVDC can connect asynchronous systems (grids operating at different frequencies or not synchronized with each other), can

be controlled precisely and rapidly, and for very long distances — typically above roughly 400 to 600 miles for overhead lines or 30 to 50 miles for underground or submarine cables — becomes more economical than AC despite the substantial cost of the converter stations required at each end to convert between AC and DC.

In the United States, HVDC links serve several critical functions. They connect the three asynchronous interconnections — the Eastern, Western, and Texas (ERCOT) Interconnections — which, despite operating at the same nominal 60-hertz frequency, are not synchronously linked and therefore cannot exchange power over AC ties. The back-to-back HVDC converter stations at locations along the seam between the Eastern and Western Interconnections allow controlled power transfers between these independently synchronized systems. Long-distance HVDC lines, such as the Pacific DC Intertie (carrying power from the Columbia River hydroelectric system in the Pacific Northwest to Los Angeles) operate at 500 kV DC and can transfer 3,100 megawatts over approximately 850 miles.

As the grid integrates large quantities of renewable energy from resource-rich but load-poor regions — Great Plains wind, desert Southwest solar — interest in new long-distance HVDC transmission has surged. Several proposed HVDC projects aim to move renewable energy thousands of miles to major load centers, a topic with significant implications for both transmission planning policy and interregional market design.

1.3.5 Transmission Limits: Thermal, Voltage, and Stability

The power-carrying capacity of a transmission line is not a single fixed number but is constrained by three distinct physical limits, with the binding constraint depending on line length and system conditions:

Thermal limits govern short lines (up to approximately 50 to 80 miles). Current flowing through the resistance of the conductor generates heat (I^2R losses), which causes the conductor to expand and sag. If the conductor sags too much, it may violate minimum clearance distances to the ground, vegetation, or structures below, creating a fault or fire risk. The thermal rating of a line is the maximum current it can carry without exceeding its maximum allowable conductor temperature — a function not only of current but also of ambient temperature, wind speed, and solar radiation on the conductor.

Voltage limits constrain intermediate-length lines (approximately 50 to 200 miles). Reactive power losses cause voltage to drop along the length of the line. If the voltage at the receiving end falls too low, equipment cannot operate properly and voltage collapse — a cascading loss of voltage stability — may occur.

Stability limits govern long lines (above approximately 200 miles). The power transferable over an AC line is proportional to the sine of the angle difference between the sending-end and receiving-end voltages. As loading increases, this angle increases. If the angle exceeds approximately 90 degrees, the system loses synchronous stability — generators at the sending end accelerate relative to those at the receiving end, and the system separates. In practice, stability limits constrain line loading to a fraction of the theoretical maximum to maintain adequate margin against disturbances.

* * *

1.4 The Distribution "Last Mile": From Substation to End-Use

1.4.1 The Step-Down Process

If the transmission system is the highway network of the electric grid, the distribution system is the network of local roads and driveways that delivers power to its final destination. The transition from transmission to distribution occurs at distribution substations, where voltage is stepped down from transmission levels to distribution levels — typically between 4 kV and 34.5 kV, with 12.47 kV and 13.8 kV being the most common primary distribution voltages in the United States.

A distribution substation is a complex facility containing one or more power transformers (stepping from transmission voltage to distribution voltage), circuit breakers and switches for protection and switching, voltage regulators, capacitor banks for reactive power compensation, and increasingly, sophisticated monitoring and control equipment. A single substation may serve 20,000 to 100,000 customers, depending on load density and local system design.

From the substation, power flows over primary distribution feeders — overhead lines on wooden poles or underground cables — at the primary distribution voltage. Near the point of end use, smaller distribution transformers (the cylindrical devices mounted on utility poles or the green pad-mounted boxes in residential neighborhoods) step the voltage down once more to the secondary or utilization voltage: 120/240 volts for residential customers in the United States, 208/480 volts for commercial and light industrial facilities.

1.4.2 Radial Versus Network Distribution Systems

Distribution systems come in two fundamentally different topologies, each with distinct reliability and cost characteristics.

Radial systems are the predominant configuration in suburban and rural areas. In a radial system, each feeder extends outward from the substation like a branch of a tree, with lateral taps branching off to serve individual customers or clusters of customers. Power flows in one direction: from the substation outward to the loads. If a fault occurs on the feeder, all customers downstream of the fault lose power until the fault is isolated and repaired. Radial systems are simple and inexpensive but inherently less reliable because any single failure interrupts service to all downstream customers.

Utilities improve radial system reliability through several techniques: automatic reclosers (devices that briefly de-energize the line to clear temporary faults, such as tree branches brushing a conductor, before automatically restoring power), sectionalizing switches (that isolate the faulted section while

restoring service to unfaulted sections), and feeder ties (normally open connections between adjacent feeders that allow load to be transferred when one feeder is out of service).

Network systems are used in dense urban cores — the downtowns of major cities like New York, Chicago, and Boston — where the cost of interruption is extremely high and load density justifies the additional investment. In a network system, multiple feeders from one or more substations supply power to an interconnected grid of secondary cables. Network protectors — specialized circuit breakers — at each feed point automatically disconnect a faulted feeder while the remaining feeders continue to supply the network. A well-designed secondary network can lose multiple feeders simultaneously without interrupting service to any customer.

The reliability difference is dramatic. Customers on radial systems in the United States experience an average of one to two sustained interruptions per year, with average cumulative duration of two to four hours. Customers on well-maintained secondary networks may experience interruptions measured in minutes per year or less.

The electric utility industry quantifies distribution reliability using two standard metrics defined by the Institute of Electrical and Electronics Engineers (IEEE). SAIDI — the System Average Interruption Duration Index — measures the total minutes of sustained interruption experienced by the average customer over the course of a year. SAIFI — the System Average Interruption Frequency Index — measures the average number of sustained interruptions per customer per year. A "sustained" interruption is typically defined as one lasting longer than five minutes, distinguishing it from momentary events like automatic recloser operations. These metrics are reported both including and excluding "major event days" — catastrophic weather events, hurricanes, ice storms — that can dramatically skew annual averages. The distinction matters: a utility may deliver excellent day-to-day reliability (low SAIDI and SAIFI excluding major events) while still exposing its customers to extended outages during extreme weather, a pattern that has become increasingly common as climate-driven storms grow more frequent and intense.

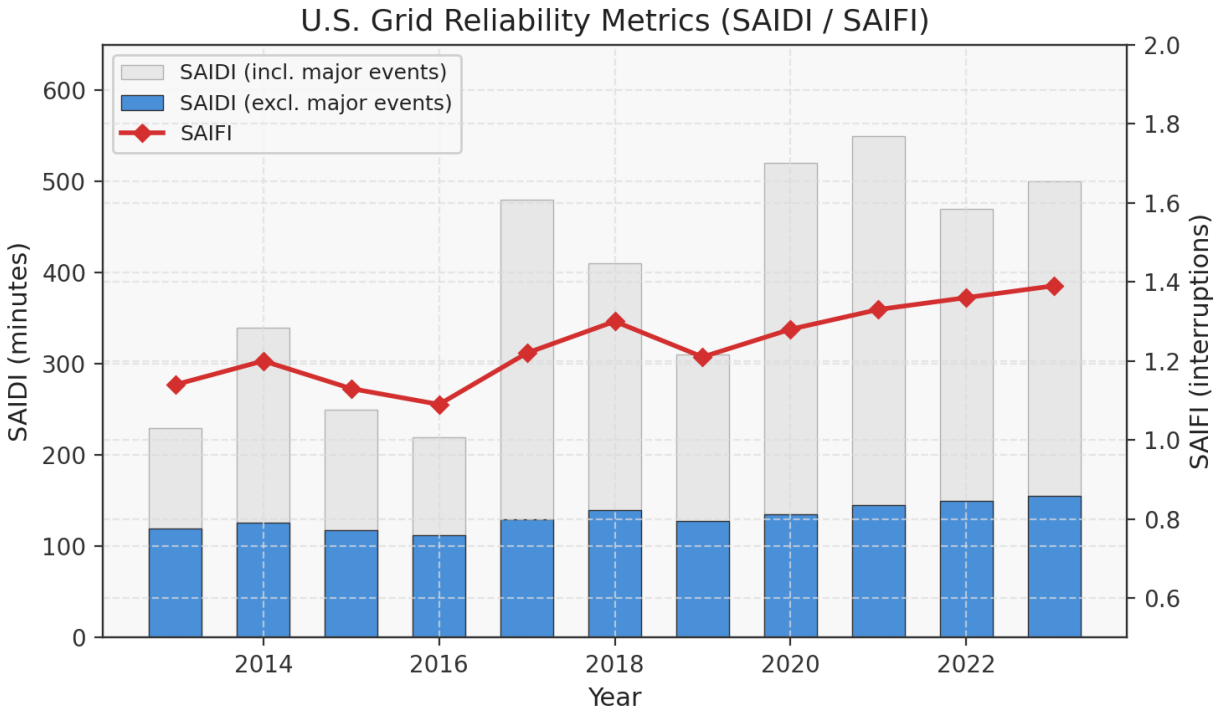

Figure 1.2: U.S. Grid Reliability Metrics — SAIDI and SAIFI Trends (Source: EIA Form EIA-861, IEEE Std 1366)

1.4.3 The Distribution System as Engineered Infrastructure

Distribution systems represent a staggering quantity of physical infrastructure. The United States has approximately 5.5 million miles of local distribution lines — roughly ten times the mileage of the transmission system. This infrastructure includes tens of millions of distribution transformers, millions of wooden poles (with an average age that in many regions exceeds forty years), and millions of miles of conductor.

Distribution has historically been the least "visible" part of the power system in policy and market discussions, despite accounting for roughly 30 to 40 percent of total electric utility investment and being the source of approximately 90 percent of all customer interruptions. Transmission garners attention because of its interstate implications and FERC jurisdiction. Generation dominates market discussions because of its competition and price formation roles. Distribution has traditionally been treated as a regulated cost-of-service function — necessary but unremarkable.

That is changing.

1.4.4 The Grid Edge: Distributed Energy Resources and Bidirectional Flow

The traditional distribution system was designed for one-way power flow: from the substation, down the feeder, through the distribution transformer, to the customer's meter. Every element — the protection schemes, the voltage regulation equipment, the conductor sizing, the transformer tap settings — was engineered on this assumption.

The proliferation of distributed energy resources (DERs) is upending this paradigm. DERs encompass a range of technologies located at or near the point of end use: rooftop solar photovoltaic systems, battery energy storage, electric vehicles (which can both consume and potentially discharge power), small-scale combined heat and power systems, controllable loads (smart thermostats, water heaters, and other devices capable of adjusting consumption in response to grid conditions), and other emerging technologies.

When a residence with rooftop solar panels generates more electricity than it consumes — as commonly occurs on a sunny midday — the excess power flows backward through the distribution transformer, onto the feeder, and potentially back toward the substation. This reverse power flow, for which the legacy distribution system was not designed, creates a suite of engineering challenges:

Voltage regulation becomes more complex because reverse power flow can cause voltage to rise above acceptable limits at points far from the substation, where voltage was traditionally at its lowest. Conventional voltage regulation devices — load tap changers on substation transformers, line voltage regulators, and switched capacitor banks — were designed for one-way flow and may not operate correctly under bidirectional conditions.

Protection coordination is complicated because fault current contributions from distributed generators can cause protective relays to misoperate — either failing to trip when they should (because fault current from the DER sustains voltage at the fault) or tripping when they should not (because the additional fault current source causes upstream devices to see current levels above their trip settings).

Thermal loading of distribution transformers and conductors may increase in unexpected patterns, as reverse power flow during peak solar production may coincide with light local load, meaning transformers originally sized for a predictable peak demand pattern now experience current flows at times and in directions for which they were not designed.

Visibility and control challenges arise because the distribution system was historically operated with minimal real-time monitoring. Utilities often had limited knowledge of conditions on individual feeders and virtually no real-time telemetry below the substation level. Managing a system with thousands or millions of active generation and storage devices requires an entirely different level of observability and control capability.

These challenges are driving massive investment in distribution system modernization — advanced metering infrastructure, distribution management systems, fault location and isolation equipment, and communication networks — and are prompting fundamental regulatory questions about the role and structure of the distribution utility. Some jurisdictions are reimagining the distribution utility as a "distribution system operator" (DSO) or "distribution system platform provider" (DSPP), responsible not only for maintaining wires and poles but for actively managing a complex, bidirectional network with thousands of active participants. These institutional innovations are examined in Part IV.

1.4.5 The Blurring Boundary

The growth of DERs is blurring the once-clear boundary between the wholesale power system (generation and transmission) and the retail distribution system. A large aggregation of residential batteries, dispatched collectively, can provide services — frequency regulation, peak demand reduction, renewable energy integration — that were once the exclusive province of utility-scale generators. A distribution feeder with high solar penetration may export power to the transmission system at certain hours, reversing the traditional hierarchy. Industrial microgrids may island themselves from the broader grid and reconnect as conditions warrant.

This blurring has profound jurisdictional implications. If an aggregation of distributed resources participates in a wholesale market, does FERC or the state commission have regulatory authority? If power flows bidirectionally across the transmission-distribution interface, how should the costs of accommodating those flows be allocated? These questions, at the intersection of physics and governance, represent some of the most contested issues in contemporary energy regulation.

* * *

Conclusion: Physics as Foundation

The four topics addressed in this chapter — synchronous frequency and the supply-demand balance, the technical taxonomy of generation, high-voltage transmission, and distribution to end-use customers — together describe the physical machine that subsequent chapters will examine through economic, regulatory, and policy lenses. Several themes from this chapter will recur throughout the book:

The *instantaneous balance requirement* — the fact that supply must equal demand at every moment, with no economically viable storage buffer at scale — drives the need for real-time markets, ancillary services, and capacity mechanisms that are examined in Part II.

The *heterogeneity of generation technologies* — their differing cost structures, operational characteristics, and environmental attributes — creates the economic dynamics of dispatch, investment, and retirement that shape market outcomes and policy interventions.

The *physical constraints of transmission* — thermal limits, stability limits, and the sheer difficulty of siting and constructing new lines — create the congestion patterns and bottlenecks that fragment markets, create locational price differences, and generate the fierce stakeholder conflicts examined in Parts II and III.

The *transformation of distribution* from passive delivery infrastructure to an active, bidirectional network is challenging a century of institutional arrangements and creating new questions about utility structure, market participation, and regulatory jurisdiction.

Understanding these physical realities is not optional for the student of electricity policy. The grid

does not care about regulatory boundaries, market designs, or political preferences. It obeys the laws of physics, and those laws impose constraints that no amount of institutional creativity can override. Every market rule, every tariff provision, every regulatory order examined in subsequent chapters is, at its foundation, an attempt to organize human institutions around the unyielding demands of the synchronous machine.

* * *

Chapter 2: The Three Great Interconnections

Introduction

Chapter 1 established the fundamental physics that govern every electric power system: the instantaneous balance of generation and load, the relationship between frequency and the rotational speed of synchronous generators, and the inescapable reality that alternating current cannot be stored at scale but must be produced and consumed in the same instant. These physical laws are not merely academic abstractions. They are the architectural constraints that dictate how the largest machine ever built by human beings—the North American power grid—came to take its present form.

If the physics of AC power systems permitted easy, lossless transmission over unlimited distances, and if synchronizing generators separated by thousands of miles posed no technical challenge, the United States might operate a single, unified grid stretching from Maine to Hawaii. It does not. Instead, the contiguous United States and parts of Canada and Mexico are served by three vast but physically distinct power systems: the Eastern Interconnection, the Western Interconnection, and the Texas Interconnection, the last of which is largely coterminous with the Electric Reliability Council of Texas, or ERCOT. Each of these interconnections is, internally, a single synchronized alternating-current machine. Every generator within a given interconnection rotates in lockstep with every other generator in that same interconnection, bound together by the physics of shared frequency. But the three interconnections are not synchronized with one another. They are, in electrical terms, islands—linked only by a small number of carefully controlled direct-current connections that permit limited power transfers between them.

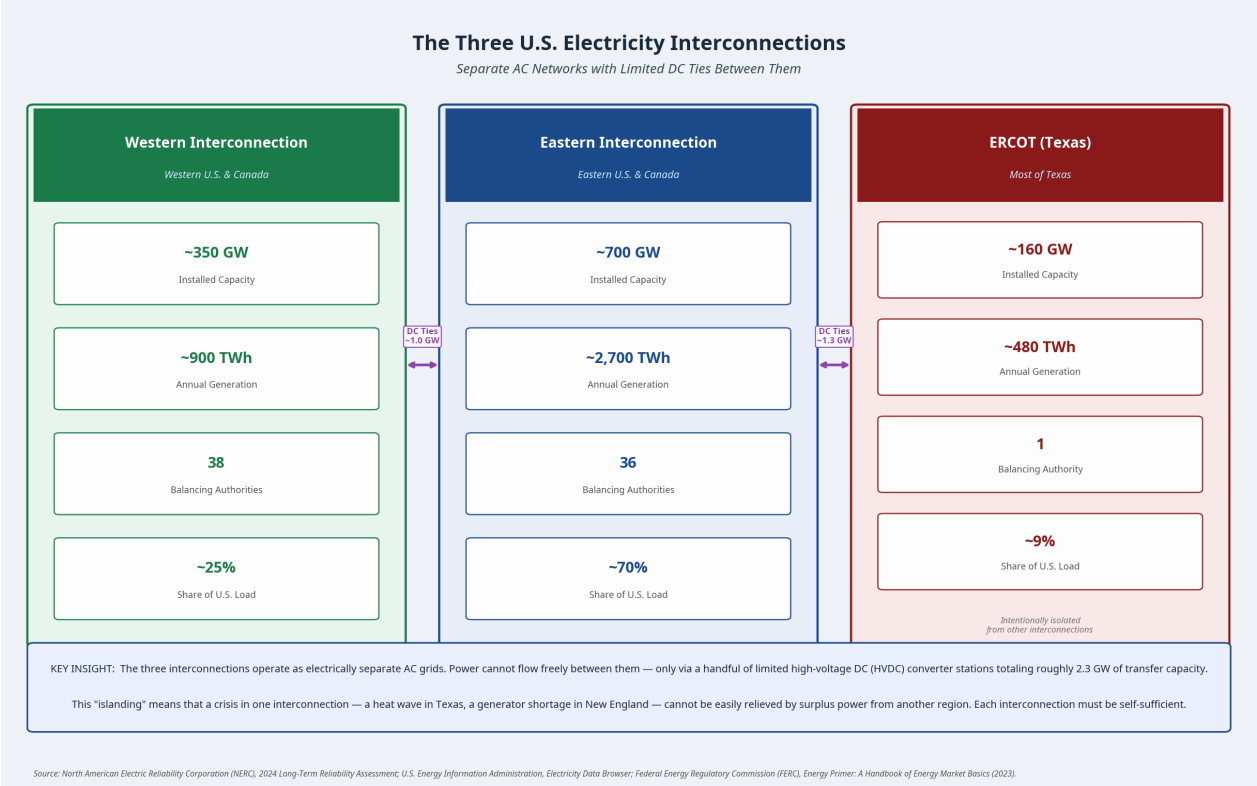

Figure 2.1: The Three U.S. Electricity Interconnections (Source: NERC, EIA, FERC)

This chapter explains how this tripartite structure came to be, why it persists, and how the systems that manage it operate. Understanding the three interconnections is essential to everything that follows in this book, because the physical boundaries of the interconnections shape markets, constrain policy, define reliability obligations, and determine who is responsible when the lights go out.

* * *

2.1 The Eastern, Western, and Texas Interconnections

2.1.1 From Isolated Systems to Interconnected Networks

The American electric power system was not designed from a master blueprint. It grew organically, utility by utility, city by city, beginning in the 1880s. Thomas Edison's Pearl Street Station in lower Manhattan, which began operation in 1882, served a territory measured in city blocks. Thousands of similar small, isolated systems followed over the next several decades. Each utility generated power and

delivered it to customers within a compact service territory. There was no particular reason, in the early years, to connect one utility's system to another's.

The incentives for interconnection emerged gradually. Utilities discovered that by linking their systems with transmission lines, they could share reserves: if one utility's generator failed, it could draw emergency power from a neighbor rather than leaving its customers in the dark. Interconnection also allowed utilities to take advantage of diversity in their load profiles. A utility whose peak demand occurred in the morning might export surplus power to a neighbor whose peak came in the afternoon. These exchanges improved reliability and reduced the total amount of generating capacity that each utility needed to build.

Through the first half of the twentieth century, these bilateral interconnections multiplied and merged. Neighboring utilities connected to one another, and those clusters connected to other clusters. The process was driven by engineering pragmatism and economic self-interest, not by any centralized plan. By the middle of the century, two large interconnected systems had taken shape in the contiguous United States: one covering most of the territory east of the Rocky Mountains, and another covering the territory to the west.

Texas followed a different path entirely, as we shall see. But in both the East and the West, the pattern was the same: bottom-up aggregation, driven by the mutual benefits of shared reserves and load diversity, producing ever-larger synchronized AC systems.

2.1.2 Why Three, and Not One?

The question naturally arises: if interconnection is beneficial, why did the process stop at three? Why not continue connecting until the entire continent operates as a single synchronized grid?

The answer lies in the physics discussed in Chapter 1. In a synchronous AC system, every generator must operate at exactly the same frequency, and the phase angles of voltage waveforms across the system must remain in a stable relationship with one another. When a disturbance occurs—a large generator trips offline, or a major transmission line faults—the effects propagate across the entire synchronized system at nearly the speed of light. Every generator in the system must respond. The larger the synchronized area, the more complex the dynamic interactions become, and the more difficult it is to maintain stability.

The Eastern and Western portions of the continent developed their networks somewhat independently during the critical decades of the early-to-mid twentieth century. By the time transmission technology might have permitted synchronous interconnection across the Rocky Mountains, the two systems had grown into enormous machines with distinct operating characteristics. Synchronizing them would have required that every generator in the East operate in precise lockstep with every generator in the West—a technical feat complicated by the vast distances involved, the different patterns of generation and load, and the risk that a disturbance in one half of the continent could cascade across the entirety of a combined system.

The Rocky Mountains themselves also played a role, though the barrier was as much economic as

topographic. Building high-voltage transmission lines across hundreds of miles of sparsely populated mountain terrain was expensive, and the electrical distance between the major load centers of the East and the West was enormous. The benefits of synchronous interconnection across such distances were modest compared to the costs and risks.

Texas, as we shall explore below, remained separate for reasons that were partly technical and partly political.

2.1.3 The Eastern Interconnection

The Eastern Interconnection is by far the largest of the three, measured by any metric one cares to apply. It stretches from the Atlantic seaboard to the eastern foothills of the Rocky Mountains, from central Canada to the Gulf Coast. It encompasses all or part of forty U.S. states, several Canadian provinces including Ontario and Quebec (which is interconnected via DC ties but operates its own synchronous system through Hydro-Québec, connected asynchronously), and a small portion of northern Mexico.

The Eastern Interconnection serves approximately 220 million people in the United States alone. Its total generating capacity exceeds 700 gigawatts, drawn from a diverse portfolio of natural gas, coal, nuclear, hydroelectric, wind, and solar resources. It contains the densest concentration of population and industry on the continent, including the great metropolitan corridors of the Northeast, the industrial Midwest, and the rapidly growing cities of the Southeast.

Within this vast territory, every synchronous generator—from the largest nuclear plant in the Tennessee Valley to a small hydroelectric station in Vermont—rotates at a speed that produces current at the same frequency, nominally 60 hertz. If the frequency at a generator in Florida is 59.98 hertz, the frequency at a generator in Minnesota is also 59.98 hertz. They are coupled, electrically and mechanically, through the web of transmission lines that binds them together. A disturbance anywhere in the Eastern Interconnection is, in principle, felt everywhere, though the magnitude of its effect attenuates with electrical distance.

2.1.4 The Western Interconnection

The Western Interconnection covers the territory from the Rocky Mountains to the Pacific Coast, including all or part of thirteen western states, the Canadian provinces of British Columbia and Alberta, and a portion of Baja California in Mexico. It serves roughly 80 million people and has a generating capacity on the order of 250 gigawatts.

The Western Interconnection is geographically vast but less densely populated than the East. Its load centers—the metropolitan areas of Los Angeles, the San Francisco Bay Area, Phoenix, Denver, Seattle, and Portland—are separated by enormous distances of desert, mountain, and rangeland. This geography has given the Western grid a distinctive character. It relies on long-distance, high-voltage transmission to a greater degree than the Eastern Interconnection, moving hydroelectric power from the Columbia River basin to California, carrying wind energy from Wyoming and Montana to coastal cities, and connecting

the solar-rich deserts of the Southwest to population centers hundreds of miles away.

The Western Interconnection is coordinated by the Western Electricity Coordinating Council (WECC), and its operating challenges are shaped by its geography: long transmission corridors that are vulnerable to wildfires, a heavy dependence on seasonal hydroelectric resources that are sensitive to drought, and the rapid growth of variable renewable generation in states with ambitious clean-energy mandates.

2.1.5 The Texas Interconnection (ERCOT)

The Texas Interconnection is unique among the three, not only in its geography but in its institutional and political character. It covers most of the state of Texas, excluding portions of the Panhandle, the eastern border region, and the El Paso area, which are electrically connected to the Eastern or Western Interconnections, respectively. The ERCOT grid serves approximately 27 million people and has a generating capacity of roughly 130 gigawatts, a figure that has grown substantially in recent years due to rapid additions of wind and solar generation.

The Texas Interconnection's independence has deep historical roots. During the massive expansion of the electric industry in the 1930s and 1940s, Texas utilities recognized that if they refrained from interconnecting with systems in other states, they could avoid the jurisdiction of the Federal Power Commission (the predecessor of the Federal Energy Regulatory Commission, or FERC). Under the Federal Power Act, federal jurisdiction over electric utilities was triggered by the transmission or sale of electricity in interstate commerce. By keeping their grid entirely within Texas, the state's utilities maintained that their operations were purely intrastate and therefore subject only to state regulation—or, in practice, to relatively little regulation at all.

This jurisdictional strategy was not the sole reason for Texas's electrical independence. The state's enormous geographic size, its abundant indigenous energy resources (first natural gas, later wind and solar), and the substantial load base of its industrial and urban economies meant that a self-contained grid was technically viable. Texas did not need to import power from neighboring states in the way that a smaller, less resource-rich state might.

The result is a power system that is, from a regulatory standpoint, unlike anything else in the United States. ERCOT operates its own wholesale electricity market, manages its own reliability standards (though it remains subject to NERC reliability standards), and functions with a degree of autonomy that is impossible for utilities embedded in the larger interconnections. This autonomy has produced both innovations—ERCOT was a pioneer in competitive retail electricity markets and in the integration of wind energy—and vulnerabilities, as the catastrophic grid failure during Winter Storm Uri in February 2021 made devastatingly clear.

2.1.6 The Boundaries Between Interconnections

The boundaries between the three interconnections are not defined by fences or walls. They are defined

by the absence of synchronous AC connections. Where AC transmission lines exist, the systems on either side are synchronized and part of the same interconnection. Where no AC path connects two systems, they belong to different interconnections, even if they are geographically adjacent.

A useful analogy is to imagine three separate clocks, each keeping its own time. Within each clock, every gear turns in coordination with every other gear. But the three clocks are not synchronized with one another. Each runs at very nearly the same rate—very nearly 60 hertz—but "very nearly" is not the same as "exactly." At any given moment, the frequency in the Eastern Interconnection might be 60.01 hertz while the frequency in the Western Interconnection is 59.99 hertz. These small differences, measured in fractions of a hertz, are sufficient to make synchronous AC interconnection between the grids impossible without the technologies discussed in the next section.

* * *

2.2 The Role of DC Ties

2.2.1 Why AC Interconnection Between the Three Grids Is Not Possible

To understand why the three interconnections cannot simply be connected with conventional AC transmission lines, one must return to the physics of synchronous systems. In an AC power system, power flows from one point to another as a function of the difference in voltage phase angles between those points. In a single synchronized grid, these phase angle differences are determined by the laws of physics and are self-correcting: if a generator begins to speed up relative to the rest of the system, the resulting change in phase angle produces forces that tend to pull it back into synchronism.

But if two systems are not synchronized—if they are operating at even slightly different frequencies—their voltage waveforms will drift in and out of phase with each other continuously. At one instant, the waveforms might align, permitting power to flow in one direction. A fraction of a second later, they might oppose each other, producing enormous currents that could damage equipment and destabilize both systems. Connecting two unsynchronized AC systems with a conventional transmission line would be roughly analogous to connecting two mechanical driveshafts that are rotating at slightly different speeds: the result would be violent and destructive.

Even if the frequencies of two interconnections could be made identical at a given instant, the phase angles of their voltage waveforms would still differ in unpredictable ways. Maintaining synchronism across the vast distances separating the interconnections—and across the boundary between systems with fundamentally different dynamic characteristics—would require a degree of control that is not practically achievable with AC technology.

2.2.2 High-Voltage Direct Current: The Bridge Between Asynchronous Worlds

The solution to this problem is high-voltage direct current (HVDC) transmission. Unlike alternating current, direct current has no frequency and no phase angle. It is simply a steady flow of electrons in one direction. By converting AC power to DC at the boundary of one interconnection, transmitting it across the boundary, and then converting it back to AC at the other side, engineers can move power between unsynchronized systems without requiring them to operate in lockstep.

The key technology is the HVDC converter station. A converter station uses power electronic devices—historically mercury-arc valves, now thyristors or insulated-gate bipolar transistors (IGBTs)—to perform the AC-to-DC and DC-to-AC conversions. At the sending end, a rectifier converts AC power into DC. At the receiving end, an inverter converts the DC power back into AC at the frequency and phase angle of the receiving system. The DC link between the two converters is electrically isolated from both AC systems, meaning that disturbances on one side do not propagate synchronously to the other.

2.2.3 Back-to-Back Converter Stations

The DC ties between the three North American interconnections are predominantly "back-to-back" HVDC converter stations. In a back-to-back configuration, the rectifier and inverter are located at the same physical site, with no intervening DC transmission line. The AC power from one interconnection enters the station, is converted to DC, and is immediately converted back to AC at the frequency and phase angle of the other interconnection. The entire process takes place within a single facility.

Back-to-back stations are used when the purpose of the HVDC link is not long-distance transmission but rather the asynchronous interconnection of two systems that are already geographically adjacent. The DC stage is simply a means of decoupling the two AC systems while permitting controlled power transfers between them.

These stations function as controllable valves. Unlike an AC transmission line, where power flow is determined by the physical characteristics of the network and cannot be directly controlled by operators, an HVDC converter station allows operators to set the direction and magnitude of power flow precisely. This controllability is one of the great advantages of DC ties: they permit power exchanges between interconnections without creating the risk of uncontrolled cascading flows.

2.2.4 Location and Capacity of the Major DC Ties

The DC ties between the Eastern and Western Interconnections are located along the seam between the two systems, running roughly along the eastern front of the Rocky Mountains. Major back-to-back facilities include those at Sidney, Montana; Rapid City, South Dakota; Stegall, Nebraska; Lamar, Colorado; and Eddy County, New Mexico, among others. Additional HVDC links connect the Eastern Interconnection to the Quebec system, which operates as a distinct synchronous area.

The ties between ERCOT and the Eastern Interconnection are located along the Texas border and include back-to-back stations at facilities such as the East and North DC ties as well as several smaller interconnections. The total transfer capacity between ERCOT and the Eastern Interconnection is on the order of 1,200 megawatts—a significant figure in absolute terms, but modest relative to ERCOT's peak demand, which can exceed 80,000 megawatts on a hot summer afternoon.

Similarly, the total transfer capacity between the Eastern and Western Interconnections through all DC ties combined is only a few thousand megawatts, a small fraction of the generating capacity of either interconnection. These limited capacities mean that the three interconnections are, for most practical purposes, self-sufficient. They cannot rely on imports from neighboring interconnections to meet significant portions of their load. Each interconnection must maintain adequate generating reserves to meet its own demand, with DC ties providing a modest supplement and a source of emergency assistance, but not a substitute for indigenous resources.

2.2.5 The Significance of Limited Interconnection

The limited transfer capacity between interconnections has profound implications for reliability and markets. When the Texas grid nearly collapsed during Winter Storm Uri in February 2021, the DC ties provided some emergency power from the Eastern Interconnection—but nowhere near enough to compensate for the massive generation shortfalls within ERCOT. The ties were operating at their maximum capacity, delivering roughly a thousand megawatts into a system that had lost tens of thousands of megawatts of generation. The physical architecture of the grid made it impossible for the rest of the country to rescue Texas, even if the generation capacity and the political will had existed to do so.

This reality—that each interconnection is largely on its own—is one of the most important facts about the American power system. It means that reliability is primarily a local and regional responsibility, that resource adequacy must be planned within each interconnection, and that the consequences of policy failures are borne principally by the population within the affected interconnection.

There have been periodic proposals to build larger HVDC links between the interconnections, motivated by the desire to share renewable energy resources across broader geographic areas, improve reliability through greater interconnection, and create more liquid wholesale electricity markets. Some of these proposals have advanced to the planning stage. The conceptual appeal is clear: a wind farm in West Texas might produce power during hours when Texas does not need it but the Eastern Interconnection does, and a large HVDC link could facilitate that exchange. But the cost of such facilities is measured in billions of dollars, the permitting and siting challenges are formidable, and the incumbent interests in each interconnection do not always favor greater integration. As of this writing, the fundamental tripartite structure of the North American grid remains intact and is likely to persist for the foreseeable future.

* * *

2.3 Balancing Authorities: The Air Traffic Controllers of the Electron World

2.3.1 The Fundamental Problem of Real-Time Balance

Chapter 1 explained that in an AC power system, supply and demand must be balanced continuously—not hourly, not every few minutes, but at every instant. If total generation exceeds total load, frequency rises. If total load exceeds total generation, frequency falls. Deviations of even a few tenths of a hertz from the nominal 60-hertz standard can cause protective relays to operate, generators to trip offline, and, in the worst case, cascading blackouts that leave millions of people without power.

Someone must be responsible for maintaining this balance. In a single small utility serving a single city, the task is conceptually straightforward: the utility's control room operators monitor load, adjust generation, and keep the system in balance. But in a vast interconnected system comprising hundreds of generators, thousands of transmission lines, and millions of customers, the task of real-time balancing cannot be performed by a single entity. It must be distributed.

This is the role of the Balancing Authority, or BA. A Balancing Authority is an entity—it may be a utility, an independent system operator, a federal power marketing agency, or some other type of organization—that is responsible for maintaining the real-time balance of generation and load within a defined geographic territory. The BA is, in effect, the air traffic controller of its portion of the grid: it monitors conditions continuously, dispatches generation to meet load, responds to contingencies, and coordinates with neighboring BAs to ensure that the interconnection as a whole remains stable.

2.3.2 What a Balancing Authority Does

The core function of a Balancing Authority is the management of generation dispatch to match load in real time. This involves several specific activities.

Load forecasting. BAs must continually forecast the demand for electricity within their territories, over time horizons ranging from minutes to days. Short-term forecasts—what load will be five minutes, fifteen minutes, or one hour from now—are essential for dispatching generation in real time. Longer-term forecasts support unit commitment decisions: which generators to start up or shut down over the coming hours and days.

Generation dispatch. Based on their load forecasts and the available generating resources, BAs issue dispatch instructions to generators within their territories, directing them to increase or decrease

output. In regions with organized wholesale markets operated by Independent System Operators (ISOs) or Regional Transmission Organizations (RTOs), this dispatch function is performed by the market operator through economic dispatch algorithms that select the least-cost combination of generators to meet load. In regions without organized markets, the BA—often a vertically integrated utility—performs dispatch based on its own fleet of generators and bilateral contracts with other generators.

Frequency regulation. BAs must ensure that sufficient generation capacity is held in reserve and responsive to automatic control signals to correct moment-to-moment imbalances between generation and load. This regulation service is provided by generators that can rapidly increase or decrease their output in response to small frequency deviations, under the direction of an Automatic Generation Control (AGC) system operated by the BA.

Contingency response. When a large generator trips offline or a major transmission line faults, the BA must have sufficient reserves to replace the lost resource quickly—typically within ten to fifteen minutes—to restore the balance of supply and demand and prevent frequency from declining to dangerous levels.

Interchange scheduling. BAs that export or import power to or from neighboring BAs must schedule these interchanges in advance and manage them in real time. Scheduled interchanges are a critical input to the BA's calculation of its own supply-demand balance: if a BA has scheduled a 500-megawatt export to its neighbor, it must generate 500 megawatts more than its own load requires.

2.3.3 Area Control Error

The primary metric by which a Balancing Authority's performance is measured is its Area Control Error, or ACE. ACE is a calculated quantity that captures the extent to which a BA is meeting its obligation to balance generation and load within its territory while honoring its scheduled interchanges with neighboring BAs.

Formally, ACE is defined as the difference between a BA's actual net interchange (the net power flowing out of the BA's territory across all of its interconnection points with neighboring BAs) and its scheduled net interchange, adjusted for a frequency bias term that accounts for the BA's share of the interconnection-wide obligation to support system frequency.

In simplified terms:

ACE = (Actual Net Interchange − Scheduled Net Interchange) − 10B(Actual Frequency − Scheduled Frequency)

where B is the BA's frequency bias setting, a parameter that reflects the sensitivity of the BA's load and generation to frequency deviations, expressed in megawatts per tenth of a hertz. The factor of 10 is a conventional scaling constant.

When ACE is zero, the BA is in perfect balance: it is generating exactly the right amount of power to meet its own load and fulfill its scheduled interchanges, and it is contributing its fair share to the interconnection-wide effort to maintain frequency at 60 hertz.

When ACE is positive, the BA is over-generating—producing more power than its load and

interchange schedules require. This surplus power flows out to the rest of the interconnection, tending to push frequency upward. When ACE is negative, the BA is under-generating, drawing power from the rest of the interconnection and tending to pull frequency downward.

NERC reliability standards require each BA to keep its ACE within specified bounds and to correct deviations within defined time periods. Two key performance standards govern this obligation. CPS1 (Control Performance Standard 1) evaluates the BA's ACE on a one-minute average basis, assessing whether the BA's ACE is helping or hurting system frequency over a twelve-month rolling period. BAAL (Balancing Authority ACE Limit) sets hard limits on ACE excursions, requiring BAs to return ACE to within bounds within a thirty-minute period. A BA that consistently fails to meet these standards is in violation of NERC reliability requirements and may face sanctions.

2.3.4 The Difference Between a Balancing Authority and a Utility

It is important to distinguish between a Balancing Authority and a utility, though the two concepts are often conflated. A utility is a business entity that generates, transmits, and/or distributes electricity and serves retail customers. A Balancing Authority is a functional role: the entity responsible for real-time balancing within a defined territory.

In many cases, a single vertically integrated utility serves as the Balancing Authority for its own service territory. The Tennessee Valley Authority, for example, functions as both a utility and a Balancing Authority. Duke Energy, Southern Company, and other large utilities in the Southeast similarly serve as BAs for their respective territories.

But in regions with organized wholesale markets, the Balancing Authority function has been consolidated into Independent System Operators or Regional Transmission Organizations. PJM Interconnection, for example, serves as the Balancing Authority for a territory spanning thirteen states and the District of Columbia, encompassing the service territories of dozens of individual utilities. The Midcontinent Independent System Operator (MISO), the California Independent System Operator (CAISO), the New York Independent System Operator (NYISO), ISO New England, the Southwest Power Pool (SPP), and ERCOT each serve as the BA for their respective footprints.

This consolidation of the BA function into larger entities has been one of the most significant structural changes in the American power industry over the past three decades. By placing the real-time balancing responsibility in the hands of a single entity that can see and optimize across a large territory, rather than leaving it distributed among dozens of smaller utilities, the organized markets have achieved significant gains in efficiency, reliability, and the integration of variable renewable resources.

2.3.5 Coordination Among Balancing Authorities

No Balancing Authority operates in isolation. Each BA is embedded in a web of relationships with its neighbors, bound by scheduled interchanges, shared reliability obligations, and the inescapable physics of the interconnected AC system. When one BA over-generates or under-generates, the effects are felt by

its neighbors in the form of unscheduled power flows.

BAs coordinate with one another through several mechanisms. Scheduled interchanges are agreed upon bilaterally or through market mechanisms and are communicated in advance. Real-time adjustments are managed through AGC systems that respond to ACE deviations. In emergency situations, BAs may request emergency energy assistance from their neighbors under established reliability protocols.

The Reliability Coordinator—a relatively recent addition to the grid's institutional architecture—provides an additional layer of oversight, monitoring conditions across a broad territory encompassing multiple BAs and directing actions when necessary to prevent reliability violations that individual BAs might not detect on their own.

2.3.6 The NERC Hierarchy: From the Continent to the Control Room

The organizational structure that governs reliability on the North American grid is hierarchical, and understanding this hierarchy is essential to understanding how the system operates.

At the top of the hierarchy sits the North American Electric Reliability Corporation, or NERC. NERC is a not-for-profit organization that serves as the Electric Reliability Organization (ERO) for North America, designated by the Federal Energy Regulatory Commission under authority granted by the Energy Policy Act of 2005. NERC develops and enforces mandatory reliability standards that apply to all users, owners, and operators of the bulk power system in the United States and Canada.

Below NERC are six Regional Entities, which are responsible for monitoring and enforcing compliance with NERC reliability standards within their respective geographic territories. These Regional Entities—which include WECC in the West, the Midwest Reliability Organization (MRO), the Northeast Power Coordinating Council (NPCC), ReliabilityFirst (RF), the Southeast Reliability Corporation (SERC), and the Texas Reliability Entity (Texas RE)—serve as the front-line compliance monitors and enforcers.

Below the Regional Entities are the Balancing Authorities themselves, along with other registered entities including Transmission Operators, Generator Operators, and Reliability Coordinators. Each of these entities has specific functional responsibilities defined by NERC's reliability standards, and each is subject to compliance monitoring and enforcement by its Regional Entity.

This hierarchy—NERC at the top, Regional Entities in the middle, and registered functional entities including BAs at the base—is the institutional framework through which reliability is managed across the continent. It is a framework that evolved in response to catastrophic events, most notably the Northeast Blackout of 2003, which demonstrated the consequences of inadequate reliability oversight and led directly to the creation of NERC's mandatory standards regime.

2.3.7 The Evolving Number of Balancing Authorities

The number of Balancing Authorities in North America has declined significantly over time. In the

mid-twentieth century, when the grid was a patchwork of hundreds of vertically integrated utilities, each utility typically served as its own Balancing Authority, or the equivalent function under whatever terminology prevailed at the time. At their peak, there were well over 100 BAs in the contiguous United States.

The formation of ISOs and RTOs in the late 1990s and 2000s consolidated many of these into larger entities. Where thirty separate utilities in the PJM footprint might once have each managed their own generation dispatch, PJM now performs this function centrally. Similar consolidation occurred in the MISO, SPP, CAISO, NYISO, and ISO New England territories.

Today, the number of BAs in North America stands at roughly 60 to 70, depending on how one counts entities in Canada and Mexico. This consolidation has generally been regarded as beneficial for reliability and efficiency, but it has not been without controversy. Smaller utilities that surrendered their BA function to an RTO sometimes chafe at the loss of operational autonomy. And the remaining unconsolidated BAs—primarily in the Southeast and parts of the West—have resisted joining organized markets, citing concerns about cost, governance, and the loss of local control.

The tension between consolidation and local autonomy is a recurring theme in the governance of the American power system, and it is one to which we shall return in later chapters.

* * *

Chapter Summary

The American power grid is not one grid but three: the Eastern Interconnection, the Western Interconnection, and the Texas Interconnection. Each is a single synchronized AC system in which every generator operates at the same frequency. The three interconnections are not synchronized with one another and are connected only by a limited number of HVDC back-to-back converter stations—DC ties—that permit controlled but modest power transfers between them.

This tripartite structure is a product of history, geography, physics, and, in the case of Texas, political calculation. The physical boundaries of the interconnections are defined not by state lines or geographic features but by the presence or absence of synchronous AC connections. The limited transfer capacity between interconnections means that each is largely self-sufficient and must maintain adequate resources to meet its own demand.

Within each interconnection, the real-time balance of generation and load is maintained by Balancing Authorities—entities that monitor conditions, dispatch generation, and manage interchanges with their neighbors. BAs operate within a hierarchical reliability framework overseen by NERC and its Regional Entities, under mandatory reliability standards that have the force of federal law.

The physical architecture described in this chapter—three interconnections, DC ties, and the BA framework—is the foundation upon which the economic and regulatory structures of the American

power system are built. The markets, the institutions, and the policies that we will examine in subsequent chapters are all shaped and constrained by these physical realities. One cannot understand American electricity markets without understanding that the Eastern Interconnection is an 800-gigawatt synchronized machine, that ERCOT is an island by choice, and that a back-to-back converter station in Nebraska is the only electrical bridge between East and West. The grid is a physical system first, an economic system second, and a political system third—and the physical architecture always has the final word.

* * *

Part II

Regulation and Ownership

Chapter 3: The Vertically Integrated Monopoly

Introduction: From Physics to Institutions

The first two chapters of this book described the physical architecture of the American power system — the generators that convert fuel into electricity, the transmission lines that carry it across vast distances, and the distribution networks that deliver it to every home, factory, and streetlight in the nation. We examined the engineering constraints that make electricity unique among commodities: it moves at the speed of light, it cannot be economically stored at scale (at least not historically), and supply must match demand in every instant or the entire system risks collapse. These physical realities are not merely technical curiosities. They shaped, and in many ways determined, the institutional structures that grew up around them.

Part II of this book turns from the physics of the grid to the political economy of the grid — the organizational forms, regulatory frameworks, and market structures through which Americans have chosen to govern their electric system. We begin with the model that dominated the twentieth century and that still serves roughly half the nation's electricity consumers: the vertically integrated, investor-owned utility operating as a regulated monopoly within an exclusive franchise territory.

This model is so familiar that it can seem natural, even inevitable. A single company generates your electricity, transmits it, distributes it to your meter, and sends you the bill. A state regulatory commission approves its rates, reviews its investments, and ensures it keeps the lights on. The company earns a regulated return on its capital; the consumer receives reliable service at a price deemed just and reasonable. For decades, this arrangement delivered declining real electricity prices, universal service, and extraordinary reliability. It was, in the judgment of many observers, among the most successful institutional arrangements in American economic history.

But the vertically integrated monopoly was not inevitable. It was constructed — assembled through entrepreneurial ambition, financial engineering, political negotiation, and economic theory. Understanding how and why it was constructed is essential to understanding the debates that later

disrupted it, and to evaluating whether it remains the right model for the challenges of the twenty-first century. This chapter tells that story in three parts: the economic theory that justified monopoly provision of electricity, the regulatory compact that governed it, and a close examination of two utilities — Duke Energy and Southern Company — that exemplify the model at its most developed.

* * *

3.1 The Natural Monopoly Theory

3.1.1 Why We Do Not Have Three Sets of Power Lines

Consider a thought experiment. Imagine that your city decided to apply the same competitive model to electricity distribution that it applies to restaurants or dry cleaners. Any company that wished to sell electricity to homes on your street would be free to do so — provided it built its own poles, strung its own wires, and installed its own transformers. Three companies might compete for your business.

The result would be absurd and immediately recognizable as wasteful. Your street would be lined with three parallel sets of utility poles. Three separate crews would dig up the pavement to lay underground conduit. Three sets of transformers would hum on three sets of poles. The capital cost of serving your neighborhood would roughly triple, yet the total amount of electricity consumed would remain the same. Each company, serving only a fraction of the customers, would spread its enormous fixed costs over a smaller base, resulting in higher per-unit costs for everyone. Far from reducing prices through competition, duplication would raise them.

This is the intuitive core of natural monopoly theory, and it explains why electricity distribution — along with water, natural gas pipelines, and local telephone service in an earlier era — has historically been organized as a monopoly rather than a competitive market.

3.1.2 Subadditivity and Economies of Scale

The formal economic theory of natural monopoly rests on the concept of subadditive cost functions. A cost function is subadditive over a relevant range of output if a single firm can produce that output at lower total cost than any combination of two or more firms. Formally, for a single-product firm, the cost function C(q) is subadditive if:

C(q) < C(q) + C(q) + ... + C(q)

for all possible divisions of total output q into portions q, q, ..., q where q + q + ... + q = q.

Subadditivity is related to, but not identical with, economies of scale. A firm exhibits economies of scale when its average cost declines as output increases — that is, when the cost of producing an additional unit falls as the firm gets larger. In industries with very high fixed costs and relatively low marginal costs, economies of scale can persist over an enormous range of output. This is precisely the cost structure of electric utilities, particularly in their transmission and distribution functions.

Building a transmission line from a power plant to a city requires an immense initial capital expenditure — acquiring rights of way, erecting towers, stringing conductors. But once the line is built, the marginal cost of transmitting an additional kilowatt-hour through it is negligible until the line approaches its thermal or stability limits. The same is true of distribution infrastructure: the poles, wires, transformers, substations, and underground conduit that constitute the local delivery network represent a massive sunk cost that does not vary meaningfully with the amount of electricity flowing through it.

The electric utility industry also benefits from economies of density — the cost per customer of building a distribution network falls as the number of customers per square mile increases. A single feeder circuit serving a dense urban neighborhood may connect hundreds of customers; the same length of wire in a rural area might serve a handful. This density effect is one reason that rural electrification proved uneconomic for private utilities and ultimately required federal intervention through the Rural Electrification Administration in 1936.

3.1.3 The Generation Question

It is important to note that the natural monopoly argument applies most forcefully to the transmission and distribution segments of the electricity value chain — the "wires" business. The case for generation as a natural monopoly was historically strong but has weakened considerably over time.

In the early decades of the industry, generating stations exhibited powerful economies of scale. Larger turbines were more thermally efficient than smaller ones. A utility that operated a single large plant could produce electricity more cheaply per kilowatt-hour than several small competitors. Through the mid-twentieth century, the optimal size of generating units grew steadily — from a few megawatts in the early 1900s to units exceeding 1,000 megawatts by the 1960s and 1970s. This trend reinforced the natural monopoly character of the entire vertically integrated chain.

But the scale economies in generation eventually exhausted themselves. By the 1970s and 1980s, the efficiency gains from building ever-larger units had plateaued, and in some cases reversed. The development of efficient combined-cycle gas turbines in the 1980s and 1990s — units that could achieve high thermal efficiency at scales of 200 to 500 megawatts — meant that relatively small, modular plants could compete on cost with the largest baseload behemoths. This technological shift was one of the key factors that later enabled the restructuring of wholesale electricity markets, as we will examine in subsequent chapters. But it did not change the fundamental monopoly economics of wires.

3.1.4 Samuel Insull and the Consolidation of the Industry

The theoretical case for natural monopoly in electricity was powerful, but theory alone did not create the vertically integrated utility. That required the vision and ambition of entrepreneurs, and none was more consequential than Samuel Insull.

Insull, born in London in 1859, came to America in 1881 to serve as personal secretary to Thomas Edison. He quickly proved himself a business genius of the first order. By 1892, he had left Edison's direct employ to become president of Chicago Edison Company, a small utility serving a fraction of Chicago's electricity consumers in a market crowded with dozens of competitors.

Insull perceived something that many of his contemporaries did not: the electric utility business had a fundamentally different cost structure than most industries, and the competitive model was therefore the wrong organizing principle. His insight was not merely about economies of scale in generation — though he was among the first to exploit those aggressively, building some of the largest generating stations of his era. More profoundly, Insull understood that the entire system — generation, transmission, and distribution — worked best as an integrated whole, and that a single company serving an entire territory could deliver electricity far more cheaply than a patchwork of competitors.

Insull pursued this vision through relentless acquisition and consolidation. He bought out competing utilities in Chicago, then expanded throughout Illinois and into neighboring states. He pioneered the concept of the exclusive franchise territory, negotiating agreements with municipal governments that granted his companies the sole right to provide electric service within defined geographic boundaries. In exchange, he accepted an obligation to serve all customers within those boundaries and submitted to rate regulation by public authorities.

Insull was also a pioneer of load management and rate design. He recognized that the cost of electricity was driven largely by the peak demand that the system had to be built to meet, not by the total energy consumed. He introduced time-differentiated rates and promotional pricing designed to fill the valleys in the load curve — encouraging off-peak consumption that would improve the utilization of his generating assets and spread fixed costs over more kilowatt-hours, thereby reducing average costs for all customers.

By the 1920s, Insull controlled a utility empire spanning multiple states, organized through an elaborate pyramid of holding companies. His Middle West Utilities Company sat atop a structure of subsidiaries and sub-subsidiaries that controlled assets worth billions of dollars with relatively modest equity at the top. The holding company structure allowed Insull — and other utility magnates of the era — to exercise control over vast enterprises while leveraging their investments to extraordinary degrees.

3.1.5 The Holding Company Era and Its Collapse

Insull was not alone. By the late 1920s, sixteen holding company groups controlled more than seventy-five percent of the nation's privately owned electric utility industry. The three largest — the Insull group, Electric Bond and Share Company (a General Electric subsidiary), and the J.P. Morgan-affiliated United Corporation — together controlled roughly forty percent.

The holding company era produced genuine efficiencies. Central management could coordinate

operations across multiple operating companies, achieve economies of scale in purchasing and engineering, and plan system expansion more rationally than isolated local utilities. But it also produced spectacular abuses. The pyramidal structures allowed holding companies to extract excessive management fees, sell overpriced services to their captive subsidiaries, manipulate inter-company transactions, and issue securities on inflated asset valuations. The operating utilities — and their ratepayers — often bore the costs.

When the stock market crashed in 1929 and the Depression deepened, the leveraged holding company pyramids began to collapse. Insull's empire was among the first and most spectacular to fall. Middle West Utilities declared bankruptcy in 1932; Insull himself fled to Europe to avoid prosecution. He was eventually extradited, tried, and acquitted on fraud charges, but his reputation was destroyed. He died in Paris in 1938 with just eighty-four cents in his pocket — a poetic if perhaps apocryphal coda to one of the most remarkable careers in American business history.

The collapse of the utility holding companies helped precipitate a fundamental restructuring of the industry's regulatory framework. The Public Utility Holding Company Act of 1935, part of President Roosevelt's New Deal, required the breakup of the interstate holding company pyramids and limited utility holding companies to single, integrated systems serving contiguous geographic areas. The Federal Power Act of the same year established federal jurisdiction over interstate wholesale electricity sales and transmission, creating the framework that would eventually give rise to the Federal Energy Regulatory Commission. Together, these statutes created the basic regulatory architecture that would govern the American electric utility industry for the next six decades. The era of freewheeling, lightly regulated holding company empires was over. In its place emerged the regulated, vertically integrated monopoly utility — operating within a defined territory, under the supervision of a state public utility commission, and subject to the regulatory compact that is the subject of the next section.

* * *

3.2 The Regulatory Compact

3.2.1 The Bargain

The regulatory compact that governs vertically integrated electric utilities is not a formal contract. No single document sets forth its terms. It is, rather, a set of interlocking legal doctrines, statutory provisions, regulatory practices, and mutual understandings that together define the relationship between the utility, its regulators, and its customers. But its essential terms can be stated simply.

The utility receives an exclusive franchise to provide electric service within a defined geographic territory. No competitor may build duplicative infrastructure to serve customers in that territory. The

utility is thus protected from competition — shielded from the market forces that discipline firms in ordinary industries.

In exchange, the utility accepts two fundamental obligations. First, it has a duty to serve — an obligation to provide electric service to any customer within its territory who requests it, at rates that are just, reasonable, and non-discriminatory. The utility cannot refuse service to customers it finds unprofitable. It cannot charge different prices to similarly situated customers. It must extend its system to reach new developments, even when the cost of doing so is high relative to the revenue the new customers will generate.

Second, the utility submits to rate regulation by a state public utility commission. It cannot charge whatever the market will bear. Its prices are set through a formal administrative process — the rate case — that is designed to allow the utility to recover its costs and earn a reasonable return on its invested capital, but no more.

This is, in essence, a grand bargain: monopoly protection and a guaranteed return in exchange for universal service and price regulation. It is a substitute for the discipline of the market — an attempt to replicate, through administrative process, the outcomes that competition would theoretically produce in an industry where competition itself is wasteful.

3.2.2 Cost-of-Service Ratemaking and the Revenue Requirement

The mechanism through which regulators set electricity prices is called cost-of-service ratemaking, and its centerpiece is the calculation of the utility's revenue requirement — the total amount of revenue the utility must collect from its customers to cover its costs and earn a fair return.

The revenue requirement formula can be expressed as:

Revenue Requirement = Operating Expenses + Depreciation + Taxes + (Rate Base × Rate of Return)

Each element of this formula warrants examination.

Operating expenses include the costs of fuel, purchased power, labor, maintenance, materials, and all other expenditures necessary to run the utility's system on a day-to-day basis. These are recovered dollar for dollar — the utility passes them through to customers without markup.

Depreciation represents the recovery of the utility's capital investment over the useful life of its assets. A generating station with a forty-year expected life would be depreciated over that period, with one-fortieth of its original cost recovered from ratepayers each year. Depreciation is an accounting mechanism that allows the utility to recover its invested capital over time while reflecting the diminishing value of aging assets.

Taxes include federal and state income taxes, property taxes, payroll taxes, and various other levies. They are passed through as a cost of service.

Rate base is perhaps the most important and contested element of the formula. The rate base is the total value of the utility's capital investments — its generating stations, transmission lines, distribution systems, substations, and other long-lived assets — that are deemed to be "used and useful" in providing

service to ratepayers. The rate base is typically calculated at the original cost of the assets (what the utility actually paid for them), less accumulated depreciation. The rate base also includes an allowance for working capital and, in some jurisdictions, construction work in progress.

Rate of return is the percentage return that the utility is authorized to earn on its rate base. The rate of return is typically set to reflect the utility's weighted average cost of capital — a blend of the interest rate on its debt and the authorized return on its equity, weighted by the proportions of debt and equity in its capital structure. The authorized return on equity — typically in the range of nine to eleven percent in recent decades — is the most closely watched and vigorously contested figure in any rate case. It must be high enough to attract investors and allow the utility to raise capital on reasonable terms, but not so high that ratepayers are paying excessive profits.

3.2.3 The Rate Case Process

The rate case is the formal proceeding through which a utility's revenue requirement and rates are established. It is a quasi-judicial process, conducted before the state public utility commission, involving sworn testimony, cross-examination, expert witnesses, and extensive documentary evidence. A major rate case can take twelve to eighteen months and generate tens of thousands of pages of testimony and exhibits.

The process typically begins when the utility files an application with the commission, asserting that its current rates are insufficient to allow it to earn its authorized return. The filing includes detailed cost studies, load forecasts, and proposed rate schedules. Intervenors — consumer advocates, industrial customer groups, environmental organizations, and other interested parties — file responsive testimony challenging the utility's claims. Commission staff conducts its own independent analysis.

The central analytical tool of the rate case is the test year — a twelve-month period whose costs and revenues serve as the basis for setting future rates. Historically, most commissions used a historical test year — typically the most recent twelve months for which audited data were available. The utility's actual costs during the test year, adjusted for known and measurable changes, established the baseline for the revenue requirement.

Some commissions have moved toward forecast or future test years, which project costs and revenues for the period when new rates will actually be in effect. Future test years are more responsive to changing conditions but introduce greater uncertainty and more opportunities for dispute. The choice between historical and future test years has significant practical consequences, particularly in periods of rapid cost change.

A critical component of the rate case is the prudency review, in which the commission evaluates whether the utility's capital investments and operating decisions were prudent at the time they were made. A utility that builds a power plant at excessive cost, or that incurs operating expenses through mismanagement, may find those costs disallowed — excluded from the rate base or from recoverable expenses. The prudency standard is typically defined as what a reasonable utility manager would have done under the circumstances known or knowable at the time the decision was made. It is not a hindsight

standard; a decision that turns out badly is not necessarily imprudent if it was reasonable when made.

3.2.4 Rate Design and Customer Classes

Once the commission has determined the utility's total revenue requirement, it must decide how to allocate that requirement among different classes of customers and how to structure the rates that each class will pay. This is the problem of rate design, and it involves both economic analysis and policy judgment.

Utilities typically serve three broad categories of customers: residential, commercial, and industrial. These classes differ in their load characteristics, their cost of service, and their sensitivity to price.

Industrial customers tend to have large, relatively stable loads with high load factors — that is, their peak demand is closer to their average demand. Because they impose relatively predictable demands on the system and can often be served at higher voltages (reducing distribution costs), the cost of serving them per kilowatt-hour is typically lower than for other classes. They also tend to be more price-sensitive and more likely to have alternatives, such as self-generation or relocation to a lower-cost jurisdiction.

Residential customers tend to have smaller, more variable loads with lower load factors. They impose peak demands on the system during summer afternoons (air conditioning) or winter mornings and evenings (heating and lighting) that drive the utility's need for generating capacity, but their average consumption may be much lower than their peak. The cost of serving them, per kilowatt-hour, is typically higher than for industrial customers because of the extensive distribution infrastructure required to reach individual homes and the peaky nature of their demand.

Commercial customers fall between these extremes. A large office building or shopping center consumes far more electricity than a single residence but operates with a more predictable daily pattern — lights and HVAC systems that ramp up in the morning and ramp down in the evening — and is typically served at a medium voltage level that requires less distribution infrastructure per kilowatt-hour than residential service.

The question of how to allocate costs among these classes involves both cost causation principles — each class should pay the costs it causes — and policy considerations, including affordability for low-income residential customers, economic development, and the administrative simplicity of rate structures. The tension between cost-based rates and policy-based rates is a perennial source of regulatory controversy.

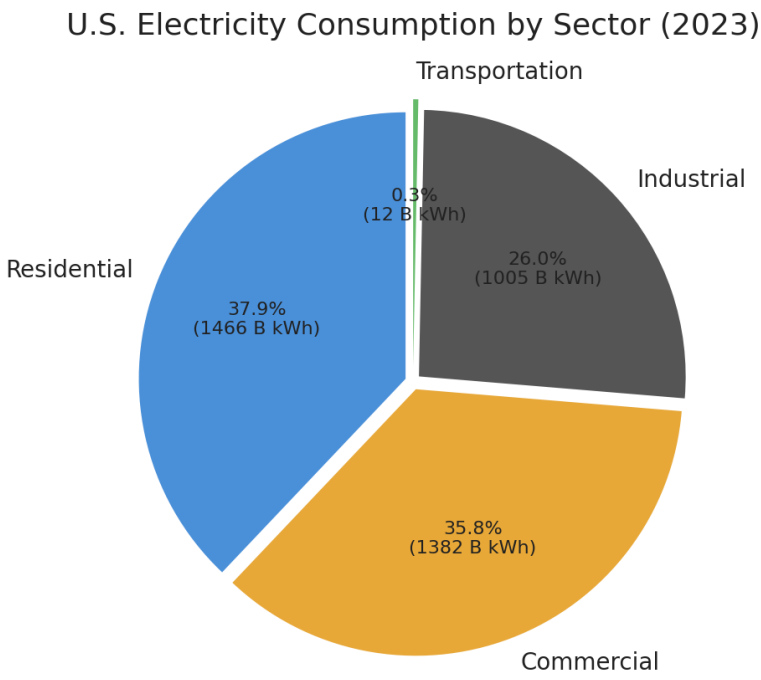

Figure 3.1: U.S. Electricity Consumption by Sector, 2023 (Source: EIA Electric Power Annual)

The Anatomy of an Electricity Bill

The electricity rate that a customer pays is not a single number. It is a composite of several distinct charges, each reflecting a different component of the cost of delivering electricity from generator to outlet.

The **customer charge** is a fixed monthly fee, typically ranging from $5 to $15 for residential customers, that recovers the cost of maintaining the customer's connection to the grid — metering, billing, and basic distribution infrastructure — regardless of how much electricity the customer consumes.

The **energy charge** is the volumetric rate, expressed in cents per kilowatt-hour, that the customer pays for each unit of electricity consumed. For a typical residential customer in the United States, the all-in energy charge ranges from roughly 10 to 25 cents per kilowatt-hour, depending on geography, utility, and rate structure. This wide range — a factor of 2.5 from cheapest to most expensive — reflects differences in fuel mix, legacy infrastructure costs, state policy choices, and climate-driven demand patterns. The regional prices chart in Figure 3.2 illustrates this geographic variation.

The **demand charge**, which is the defining feature of most commercial and industrial rate schedules, is expressed in dollars per kilowatt of peak demand during the billing period. Where the energy charge asks *how much* electricity the customer consumed, the demand charge asks *how fast* the customer consumed it at the moment of peak draw. A factory that uses 100,000 kilowatt-hours per month

with a steady 140-kilowatt load pays a very different demand charge than a factory that uses the same total energy but with sharp peaks of 500 kilowatts during shift changes. The demand charge exists because the utility must build and maintain generation, transmission, and distribution capacity sufficient to serve the customer's peak load, even if that peak occurs for only a few minutes per month. For large commercial and industrial customers, demand charges can constitute 30 to 50 percent or more of the total bill — a fact that drives significant investment in load management, power factor correction, and on-site generation.

Residential customers, by contrast, typically do not face explicit demand charges. Their contribution to system peak demand is instead recovered through higher volumetric energy charges, a simplification that sacrifices cost precision for billing simplicity. This cross-subsidy — residential customers who consume at off-peak times effectively subsidize those who drive the system peak — is one of the perennial tensions in rate design.

Beyond these core charges, most electricity bills include a constellation of riders, surcharges, and adjustments: fuel adjustment clauses that pass through changes in the utility's fuel costs between rate cases, renewable energy surcharges that recover the cost of complying with state renewable portfolio standards, infrastructure trackers that recover specific capital investments on an accelerated schedule, and various state and local taxes. The sum of these components transforms the wholesale cost of electricity — which might be $30 to $50 per megawatt-hour in an organized market — into the retail rate that appears on the customer's bill.

From Wholesale to Retail: The Price Bridge

The gap between wholesale and retail electricity prices is substantial, and understanding its composition is essential to understanding the economics of the electric power system.

In regions with organized wholesale markets, the cost of energy — the actual kilowatt-hours generated by power plants — typically represents only 30 to 50 percent of the retail price. The remainder consists of transmission charges (recovering the cost of the high-voltage network that moves power from generators to load centers), distribution charges (recovering the cost of the local infrastructure — substations, transformers, poles, and wires — that delivers power to individual premises), capacity charges (recovering the cost of maintaining enough generating capacity to serve peak demand with a reserve margin, whether through RTO-administered capacity markets or utility-owned generation), and the various riders, taxes, and surcharges described above.

The precise allocation varies dramatically by utility and jurisdiction. In New England, where aging transmission infrastructure and pipeline-constrained natural gas supply drive high energy and capacity costs, the wholesale energy component might be only 35 percent of the retail rate. In the Pacific Northwest, where abundant low-cost hydroelectric generation keeps wholesale prices low but extensive distribution systems must still be maintained, the distribution component may dominate. In Texas, where the energy-only ERCOT market has no separate capacity payments, wholesale energy costs (including scarcity pricing episodes) represent a larger share of the retail price, but the transmission and distribution charges are layered on by the regulated utilities that own the wires.

This wholesale-to-retail price bridge explains a phenomenon that often puzzles casual observers:

retail electricity prices in the United States have remained remarkably stable in real terms over decades, even as wholesale market prices have fluctuated dramatically. The wholesale energy component is only one input to the retail rate. When natural gas prices collapsed in the 2010s, wholesale electricity prices fell substantially, but retail rates declined much less — because the non-energy components of the bill, driven by infrastructure investment and policy mandates, continued to rise.

Modern Rate Design: Aligning Price with Cost

The traditional rate structure — a flat per-kilowatt-hour charge that is the same regardless of when the electricity is consumed — is an increasingly poor reflection of the actual cost of producing and delivering electricity. The cost of generating electricity varies enormously over the course of a day: cheap at 3:00 a.m. when demand is low and baseload generators are sufficient, expensive at 5:00 p.m. on a hot summer afternoon when every available generator is running and the system is near its peak.

Tiered rates (also called inclining block rates) charge a higher per-kilowatt-hour price as a customer's consumption increases within a billing period. California has been the most aggressive adopter, with residential rates that can exceed 50 cents per kilowatt-hour in the highest tier. Tiered rates are designed to promote conservation and protect low-usage customers from high bills, but they bear no relationship to *when* the electricity is consumed and can perversely penalize large households or customers who have electrified their heating and transportation — precisely the behavior that decarbonization policy seeks to encourage.

Time-of-use (TOU) rates charge different prices at different times of day, typically distinguishing between peak, shoulder, and off-peak periods. A TOU rate might charge 30 cents per kilowatt-hour between 4:00 and 9:00 p.m. (when system demand peaks), 15 cents during shoulder hours, and 8 cents overnight. TOU rates send a price signal that encourages customers to shift flexible loads — running the dishwasher, charging the electric vehicle, pre-cooling the house — to off-peak hours when the system has surplus capacity and generation costs are low. California, Arizona, and several other states now default residential customers to TOU rates, and the rapid growth of behind-the-meter battery storage is partly a response to the arbitrage opportunity that TOU rate differentials create.

Critical peak pricing and **variable peak pricing** go further, imposing very high rates — sometimes exceeding $1 per kilowatt-hour — during a limited number of system stress events each year, with advance notice to customers. These rates are designed to elicit demand reduction precisely when the system is most constrained and the marginal cost of generation is highest.

Real-time pricing, in which the retail rate tracks the wholesale market price on an hourly or sub-hourly basis, represents the most economically efficient rate design — and the most demanding for customers to manage. Few residential customers are willing to monitor wholesale electricity prices continuously, but for large commercial and industrial customers with sophisticated energy management systems, real-time pricing can deliver substantial savings. The Griddy experience in ERCOT during Winter Storm Uri — in which retail customers on a real-time pricing plan received bills of $10,000 or more for a single week of consumption — illustrated both the theoretical efficiency and the practical risks of exposing retail customers to wholesale price volatility without adequate hedging or price caps.

The evolution of rate design is inextricable from the broader transformation of the grid. As

distributed energy resources proliferate, as electric vehicles add new and potentially flexible loads, and as the grid's generation mix shifts toward resources with zero marginal cost but variable output, the question of how to price electricity — and the gap between wholesale cost and retail price — becomes not just an economic puzzle but a matter of system reliability and the pace of the energy transition.

3.2.5 The Averch-Johnson Effect

In 1962, economists Harvey Averch and Leland Johnson published a seminal paper identifying a systematic distortion embedded in the rate-of-return regulatory model. Their insight, which came to be known as the Averch-Johnson effect, was elegant and troubling.

Under cost-of-service regulation, the utility earns a return on its rate base — its invested capital. If the authorized rate of return exceeds the utility's actual cost of capital (as it typically does, at least for equity), then each dollar added to the rate base generates a profit for the utility's shareholders. The utility therefore has a systematic incentive to expand its rate base — to favor capital-intensive solutions over less capital-intensive alternatives, to build rather than buy, to gold-plate its facilities, and to resist innovations that would reduce the need for capital investment.

Consider a concrete example. Suppose a utility can meet growing demand either by building a new power plant (a capital-intensive solution that adds to the rate base) or by contracting with an independent generator for purchased power (an operating expense that does not add to the rate base). Even if the purchased power option is cheaper for ratepayers, the utility's shareholders benefit more from building the plant, because the plant goes into the rate base and earns a return. The utility's financial incentive is misaligned with the ratepayer's interest.

The Averch-Johnson effect does not mean that utilities deliberately waste money or that regulators are powerless to prevent it. Prudency reviews, independent cost audits, and competitive procurement requirements all serve as checks on gold-plating. But the structural incentive is real, and it has influenced utility behavior in ways both subtle and profound. Critics of the traditional regulatory model have argued that the Averch-Johnson effect contributed to the massive over-building of nuclear power plants in the 1970s and 1980s — projects that, whatever their engineering merits, resulted in tens of billions of dollars in costs that ratepayers are still recovering.

The Averch-Johnson effect also has implications for the modern grid. When a utility can choose between a traditional capital-intensive solution (say, building a new transmission line to relieve congestion) and a non-wires alternative (such as targeted energy efficiency, demand response, or distributed energy resources), the regulatory incentive still tilts toward the capital project. Reforming this incentive structure — through performance-based regulation, decoupling, or other mechanisms — is one of the central challenges of contemporary utility regulation, a topic we will take up in later chapters.

* * *

3.3 Case Study: Duke Energy and Southern Company

3.3.1 Two Giants of the Vertically Integrated Model

To understand the vertically integrated monopoly not as an abstraction but as a living institution, it is instructive to examine two of its most prominent exemplars: Duke Energy and Southern Company. These two utilities, both headquartered in the southeastern United States, are among the largest electric utilities in the nation. Together, they serve more than fifteen million customers across a swath of territory stretching from Indiana to Florida. They are products of the regulatory and economic environment described in the preceding sections, and they illustrate both the strengths and the tensions of the traditional utility model.

3.3.2 Duke Energy: Profile and History

Duke Energy, headquartered in Charlotte, North Carolina, traces its origins to the Catawba Power Company, founded in 1899 to supply electricity to textile mills along the Catawba River in the Carolina Piedmont. The company grew through the early twentieth century under the leadership of James Buchanan Duke, the tobacco and hydroelectric power magnate who consolidated electric utilities across the Carolinas and invested heavily in hydroelectric generation. Duke Power, as it was known for most of the twentieth century, became the dominant electric utility in the western Carolinas, building a substantial portfolio of hydroelectric and coal-fired generation and operating one of the most reliable distribution systems in the country.

In the early twenty-first century, Duke Power merged with Cinergy Corporation (an Ohio and Indiana utility) in 2006, creating a multi-state utility holding company that it named Duke Energy. A subsequent merger with Progress Energy in 2012 — at the time the largest utility merger in American history — further expanded Duke's footprint to include eastern North Carolina, South Carolina, and Florida. Today, Duke Energy serves approximately eight million customers across six states: North Carolina, South Carolina, Florida, Indiana, Ohio, and Kentucky.

Duke's generation portfolio reflects the history and geography of its service territory. In the Carolinas, the company operates a significant fleet of nuclear generating stations — including the McGuire and Catawba nuclear plants near Charlotte and the Oconee plant in South Carolina — alongside coal, natural gas, and hydroelectric generation. In Indiana, Duke inherited a heavily coal-dependent portfolio from Cinergy, which the company has been gradually transitioning toward natural gas and renewables. In Florida, the generation mix is dominated by natural gas.

Duke operates under the regulatory jurisdiction of multiple state commissions, including the North Carolina Utilities Commission, the Public Service Commission of South Carolina, the Florida Public Service Commission, and the Indiana Utility Regulatory Commission. Each commission has its own

regulatory traditions, rate-setting methodologies, and policy priorities, creating a complex patchwork of regulatory requirements that the company must navigate.

3.3.3 Southern Company: Profile and History

Southern Company, headquartered in Atlanta, Georgia, has a history that runs parallel to Duke's but in the Deep South. The company was formed in 1945 as a successor to the southeastern operations of the old Commonwealth and Southern Corporation — itself a product of the holding company era that the Public Utility Holding Company Act of 1935 was designed to reform. Southern Company operates through four major regulated utility subsidiaries: Georgia Power, Alabama Power, Mississippi Power, and (since its 2016 merger with AGL Resources) Southern Company Gas. The company also operates Gulf Power in the Florida panhandle, which was merged into Florida Power and Light's parent in 2019 but whose territory illustrates the company's historical reach.

Southern Company serves approximately nine million electric and gas customers, making it one of the largest utility holding companies in the United States. Its electric service territory encompasses most of Georgia and Alabama, along with portions of Mississippi and Florida — a vast geographic area with a diverse mix of urban, suburban, and rural customers.

Southern Company's generation portfolio has historically been dominated by coal and nuclear power. The company operates two major nuclear stations — Plant Hatch and Plant Vogtle in Georgia — along with a large fleet of coal and natural gas plants. Southern Company became the subject of intense national attention with its decision to build two new nuclear units at the Vogtle site — Vogtle Units 3 and 4 — the first new nuclear construction in the United States in a generation. The project, originally estimated to cost approximately fourteen billion dollars, experienced massive cost overruns and schedule delays, ultimately costing more than thirty-five billion dollars and entering service years behind schedule. Unit 3 began commercial operation in 2023 and Unit 4 in 2024. The Vogtle expansion became a cautionary tale about the risks of large capital projects under cost-of-service regulation and a focal point of debates about the Averch-Johnson effect and the adequacy of regulatory oversight.

3.3.4 The Regulatory Environment: Why the Southeast Remains Vertically Integrated

One of the most striking features of the American electric utility landscape is the geographic divide between regions that restructured their electricity markets in the late 1990s and early 2000s — introducing wholesale competition and, in some cases, retail choice — and regions that retained the traditional vertically integrated model. The Southeast is the heartland of the traditional model. North Carolina, South Carolina, Georgia, Alabama, Mississippi, and Florida all retain vertically integrated utilities operating under cost-of-service regulation without retail choice or organized wholesale markets.

This is not accidental. Several factors explain why the Southeast resisted the restructuring wave that

swept the Northeast, Mid-Atlantic, and parts of the Midwest and Texas.

First, electricity prices in the Southeast were — and remain — relatively low by national standards. The primary driver of restructuring in states like California, New York, and New England was high electricity prices. Ratepayers and policymakers in those regions believed that competition would drive down costs. In the Southeast, where coal was cheap and abundant, nuclear plants had been built at relatively reasonable cost (at least compared to the spectacular cost overruns experienced at some northeastern nuclear projects in the 1970s and 1980s), and the overall cost structure was favorable, the political impetus for restructuring was weak. It is difficult to build a coalition for radical change when the existing system is delivering affordable, reliable service.

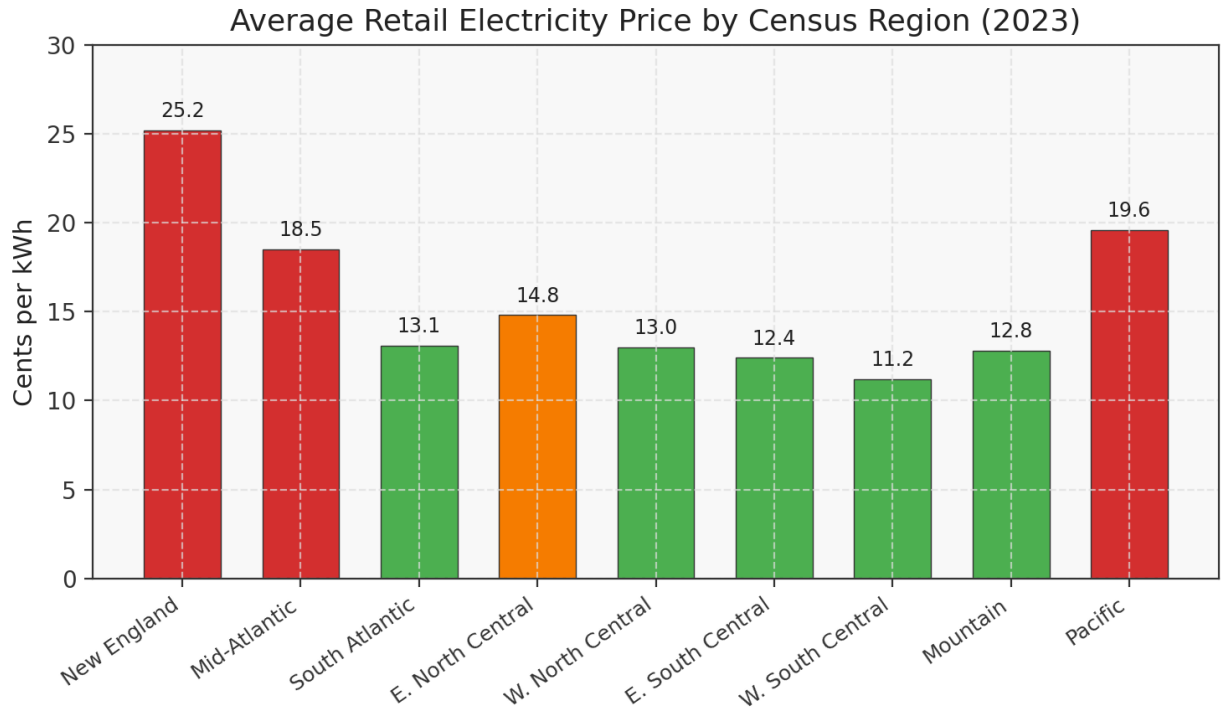

Figure 3.2: Average Retail Electricity Price by Census Region, 2023 (Source: EIA Electric Power Annual)

Second, the political culture of the southeastern states tends to favor stable, established institutions and close relationships between regulated industries and state government. The public utility commissions in Georgia, Alabama, and the Carolinas have historically maintained cooperative — critics would say cozy — relationships with their regulated utilities. State legislators in these states have generally been receptive to the arguments of utilities that vertical integration and cost-of-service regulation serve the public interest, and skeptical of the claims of restructuring advocates. The utilities themselves have been sophisticated and effective political actors, investing heavily in government relations, campaign contributions, and public affairs.

Third, the Southeast lacks the organized regional transmission organizations (RTOs) and independent system operators (ISOs) that facilitate competitive wholesale markets in other parts of the

country. While most of the nation's electricity is now traded through organized wholesale markets administered by entities like PJM, MISO, SPP, ERCOT, ISO-NE, and NYISO, the Southeast has resisted the formation of such organizations. Duke Energy and Southern Company dispatch their own generation and manage their own transmission systems, maintaining the vertically integrated model in which the same entity that owns the generation also controls access to the transmission grid. Proposals to form a southeastern RTO or to require these utilities to join an existing one have been repeatedly considered and repeatedly rejected.

3.3.5 Adaptation and Resistance

While Duke Energy and Southern Company have preserved the fundamental structure of the vertically integrated model, they have not been entirely static. Both companies have made significant investments in renewable energy, driven by declining costs, customer demand, and state policy requirements. Duke Energy has announced ambitious clean energy transition plans for its Carolinas service territories, including substantial investments in solar generation, battery storage, and grid modernization, though these plans have generated intense regulatory and political debate about pace, cost, and reliability. Southern Company has similarly invested in solar and wind resources, though its generation portfolio remains more carbon-intensive than the national average.

Both companies have also embraced grid modernization investments — advanced metering infrastructure, distribution automation, and grid-edge technologies — which, not coincidentally, add to the rate base and generate returns for shareholders. Critics have pointed to this pattern as a contemporary manifestation of the Averch-Johnson effect: utilities enthusiastically investing in capital-intensive grid modernization while showing less enthusiasm for non-wires alternatives, demand-side management, or third-party distributed energy resources that might reduce the need for utility-owned capital.

The Vogtle nuclear project provides the most dramatic illustration of the tensions inherent in the vertically integrated model. Southern Company's subsidiary Georgia Power undertook the project with the approval of the Georgia Public Service Commission, which allowed the company to collect construction financing costs from ratepayers during the construction period through a mechanism known as construction work in progress, or CWIP, in rate base. This mechanism shifted much of the construction risk to ratepayers during the long and troubled construction period. As costs escalated far beyond original estimates, the commission conducted periodic reviews but ultimately allowed recovery of most of the costs, concluding that completing the project was preferable to abandoning it after billions had already been spent. Whether this outcome represents the regulatory compact working as intended — the commission exercising oversight while ultimately allowing recovery of prudently incurred costs — or a failure of regulatory discipline that allowed ratepayers to bear the consequences of inadequate project management, is a question that will be debated for decades.

Duke Energy has had its own controversies, most notably the massive coal ash spill at its Dan River Steam Station in 2014 — one of the largest coal ash spills in American history — which raised questions about the company's environmental stewardship and its regulators' oversight. The subsequent coal ash

cleanup costs, running into billions of dollars, became the subject of contentious rate cases in which the North Carolina Utilities Commission had to determine how much of the cleanup cost should be borne by ratepayers and how much by shareholders. The commission ultimately split the costs, disallowing a portion as imprudent — a relatively rare exercise of regulatory authority that illustrated both the power and the limits of the prudency review process.

3.3.6 The Persistence of the Model

Despite these challenges, the vertically integrated model has shown remarkable staying power in the Southeast. Duke Energy and Southern Company remain among the most financially successful utilities in the country, delivering consistent returns to shareholders while maintaining relatively affordable electricity rates for their customers. Their stock prices reflect investor confidence in the stability and predictability of the cost-of-service regulatory model — a model that, whatever its theoretical inefficiencies, offers a degree of earnings visibility that competitive market participants cannot match.

The persistence of the model is also self-reinforcing. Because the Southeast lacks organized wholesale markets, there is no ready alternative to utility-owned generation. Because the utilities own both generation and transmission, potential competitors face significant barriers to entry. Because the regulatory commissions and state legislatures have long-standing relationships with the utilities, proposals for structural change face well-organized opposition. And because electricity prices remain relatively low, the political constituency for change remains small.

Whether this equilibrium will hold in the face of accelerating technological change — the plummeting costs of solar and wind generation, the emergence of battery storage, the electrification of transportation and buildings, and the growing sophistication of distributed energy resources — is one of the central questions of contemporary energy policy. The vertically integrated model was built for a world of large, central-station power plants, one-directional power flows, and passive consumers. The emerging energy landscape looks very different. How Duke Energy, Southern Company, and their peers adapt to this new reality — and how their regulators reshape the regulatory compact to accommodate it — will determine whether the vertically integrated monopoly remains a viable model for the twenty-first century or becomes a relic of the twentieth.

* * *

Chapter Summary

The vertically integrated electric utility monopoly is one of the most consequential institutional innovations in American economic history. It arose from the interaction of physical necessity, economic logic, entrepreneurial ambition, and political choice. The natural monopoly characteristics of

transmission and distribution networks made competitive provision of electricity wasteful; the regulatory compact substituted administrative oversight for market discipline; and the resulting model delivered decades of declining costs, expanding service, and extraordinary reliability.

But the model is not without its pathologies. The Averch-Johnson effect creates systematic incentives for over-investment. The rate case process is costly, slow, and imperfect. The close relationship between utilities and their regulators creates risks of capture and complacency. And the model's fundamental assumption — that a single entity should own and control the entire chain of electricity production, delivery, and sale — is increasingly challenged by technological and economic developments that favor distributed, competitive, and customer-driven alternatives.

The next chapter examines the forces that disrupted this model beginning in the 1970s — the energy crises, the rise of independent power producers, the passage of the Public Utility Regulatory Policies Act, and the intellectual revolution in electricity market design that led to restructuring and the creation of organized wholesale markets. The vertically integrated monopoly did not disappear, but it lost its claim to inevitability. Understanding why requires first understanding what it was, how it worked, and why it succeeded for as long as it did. That has been the task of this chapter.

* * *

Chapter 4: Public Power and Federal Interventions

The story of American electrification is often told as a triumph of private enterprise — of Edison, Insull, and the great holding companies that wired the nation's cities in the early twentieth century. But this narrative, however compelling, is radically incomplete. By the early 1930s, roughly ninety percent of urban Americans enjoyed electric service, yet barely ten percent of rural households had access to electricity. The farms and small towns that fed the nation did so by kerosene lamp. The rivers that carved the continent's great valleys ran unharnessed to the sea. The gap between electrified America and dark America was not merely a technological inconvenience; it was an economic and moral chasm that would provoke one of the most consequential expansions of federal power in the nation's history.

This chapter examines the public side of the American power system — the federal agencies, public utilities, and cooperative enterprises that arose in response to the perceived failures of investor-owned utilities to deliver universal electric service. These institutions, born primarily of the New Deal era but extending their influence well into the twenty-first century, represent a fundamentally different philosophy of electricity provision: that electric power is not solely a commodity to be sold at profit, but a public resource to be developed for the broad benefit of the people. Whether one regards this philosophy as visionary statesmanship or dangerous government overreach, its institutional legacy is undeniable. Federal power marketing administrations, the Tennessee Valley Authority, the Bonneville Power Administration, and hundreds of rural electric cooperatives together serve tens of millions of Americans and control some of the most valuable generation assets in the world. Understanding how they came to exist, how they operate, and how they interact with the broader electricity market is essential to any serious study of the American grid.

* * *

I. The New Deal Era: Federal Power as National Policy

The Crisis of Rural Darkness

To understand why the federal government entered the electricity business, one must first appreciate the depth of rural America's exclusion from the electric age. In 1930, according to census data, only about ten percent of American farms had electric service. In some Southern and Western states, the figure was closer to three percent. The reasons were straightforward economics: rural customers were spread across vast distances, requiring miles of distribution line to serve a handful of homes. The revenue per mile of line was a fraction of what urban service generated. Private utilities, answerable to shareholders and governed by the obligation to earn a reasonable return on invested capital, had little financial incentive to extend service to areas where the cost of construction far exceeded any plausible revenue stream.

This was not, in the view of most utility executives, a failure of will but a reality of mathematics. The National Electric Light Association, the industry's trade group, argued strenuously that rural electrification was simply uneconomic — that farmers could not afford to pay rates sufficient to justify the capital investment required to reach them. Some utilities experimented with rural extension programs, but these were limited in scope and often required farmers to pay the full cost of line construction upfront, a prohibitive barrier for families already struggling through the agricultural depression that preceded and then merged with the Great Depression.

Critics of the private utility industry saw the matter differently. To progressives, labor leaders, and agrarian populists, the refusal to electrify rural America was evidence not of economic rationality but of monopolistic indifference. The holding company abuses documented in the Federal Trade Commission's landmark investigation of the late 1920s and early 1930s — the pyramidal corporate structures, the inflated asset valuations, the self-dealing service contracts — suggested that private utilities were more interested in extracting profit from captive urban ratepayers than in expanding the reach of electric civilization. The collapse of Samuel Insull's utility empire in 1932, which destroyed billions of dollars in investor wealth, only reinforced the narrative that the private utility model was fundamentally broken.

FDR's Vision and the Politics of Public Power

Franklin Delano Roosevelt came to the presidency in 1933 with unusually well-developed views on electric power. As Governor of New York, he had championed the development of the St. Lawrence River for public hydroelectric generation and had sparred repeatedly with private utilities over rates and service territory. Roosevelt believed that electricity was a necessity of modern life, akin to water or roads, and that government had a legitimate role in ensuring its broad availability. He also believed — and this was a point of considerable technical controversy — that public power projects could serve as a "yardstick" against which the rates and efficiency of private utilities could be measured. If a government-owned dam could generate electricity and deliver it to consumers at rates dramatically below those charged by private utilities, that fact alone would demonstrate that private rates were inflated by monopoly rents, excessive capitalization, and holding company extraction.

This yardstick theory was anathema to the private utility industry, which argued that the comparison

was inherently unfair. Federal power projects paid no taxes, borrowed at the government's cost of capital, and faced none of the regulatory and financial obligations imposed on investor-owned utilities. Any rate comparison, the industry contended, was apples to oranges. But Roosevelt was unmoved. In his view, the existing system had failed a fundamental test — it had left the majority of rural Americans in darkness — and bold federal action was both justified and necessary.

The Tennessee Valley Authority

The Tennessee Valley Authority, created by act of Congress on May 18, 1933, was among the most ambitious and controversial experiments in American governance. The TVA Act established a federally owned corporation with sweeping authority over the Tennessee River and its tributaries — a vast watershed spanning parts of seven states, encompassing some of the poorest and most underdeveloped regions of the nation. The Authority was charged not only with generating and selling electric power but with controlling floods, improving navigation, producing fertilizer, promoting reforestation, and fostering the economic development of the entire Tennessee Valley.

The power provisions of the TVA Act reflected Roosevelt's philosophy with precision. TVA was directed to give preference in the sale of its electricity to public bodies and cooperatives — municipalities, counties, and rural electric cooperatives — rather than to private utilities. This "preference power" doctrine, which would become a defining feature of all federal power programs, embodied a deliberate policy choice: federal hydroelectric resources, developed with taxpayer dollars, should benefit public and cooperative entities that would pass the low cost of that power through to consumers, rather than private utilities that might use cheap federal power to enhance shareholder returns.

TVA's early years were consumed by legal and political combat. Private utilities, led by Wendell Willkie's Commonwealth and Southern Corporation, challenged TVA's constitutionality and fought to prevent municipalities from forming public power systems to purchase TVA electricity. The Supreme Court ultimately upheld TVA's authority in *Ashwander v. Tennessee Valley Authority* (1936), though on relatively narrow grounds. Meanwhile, TVA proceeded to build a massive system of dams along the Tennessee River and its tributaries — Norris Dam, Wheeler Dam, Pickwick Landing Dam, and many others — creating a chain of reservoirs that simultaneously controlled flooding, improved navigation, and generated enormous quantities of cheap hydroelectric power.

The results were transformative. In the TVA service territory, electricity rates fell dramatically, and electric consumption per household rose far above the national average — vindicating, in the eyes of public power advocates, the thesis that demand for electricity was highly elastic and that private utilities had been suppressing consumption through excessive pricing. Rural electrification proceeded rapidly as municipalities and cooperatives formed to purchase TVA preference power. The valley's economy, long one of the most depressed in the nation, experienced measurable improvement in agricultural productivity, industrial development, and household living standards.

Over the decades that followed, TVA expanded well beyond its hydroelectric origins. During World

War II, TVA's cheap electricity powered the uranium enrichment facilities at Oak Ridge, Tennessee — one of the key installations of the Manhattan Project. In the postwar era, TVA built a fleet of coal-fired power plants to meet surging demand, and in the 1960s and 1970s, it embarked on an ambitious nuclear construction program. By the late twentieth century, TVA had become the largest public power company in the United States, with a generating capacity exceeding 30,000 megawatts and a service territory encompassing virtually all of Tennessee and portions of six surrounding states. It served not individual retail customers but rather 153 local power companies — municipal utilities and rural electric cooperatives — that in turn served approximately ten million people.

TVA's institutional structure is unique in American government. It is a federally chartered corporation governed by a board of directors appointed by the President and confirmed by the Senate. It receives no congressional appropriations for its power program; instead, it finances its operations entirely through electricity sales and bond issues. It pays no federal income taxes but makes payments in lieu of taxes to state and local governments in its service territory. It is exempt from regulation by the Federal Energy Regulatory Commission, setting its own rates subject to its statutory mandate to sell power at rates "as low as are feasible." This hybrid status — neither a traditional government agency nor a private corporation — has made TVA a perennial subject of political debate, admired by public power advocates as proof that government can efficiently manage a large power system, and criticized by free-market proponents as a subsidized competitor that distorts electricity markets.

The Bonneville Power Administration

While TVA was transforming the Southeast, a parallel story was unfolding in the Pacific Northwest. The Columbia River, the largest river by volume in western North America, represented an almost incomprehensibly vast hydroelectric resource. The Bonneville Project Act of 1937 established the Bonneville Power Administration to market the electricity generated by the Bonneville Dam, then under construction by the Army Corps of Engineers on the Columbia River approximately forty miles east of Portland, Oregon.

BPA's mandate was, in certain respects, narrower than TVA's. It was not charged with comprehensive regional development but rather with the more focused task of marketing and transmitting electricity from federal dams. Yet its impact on the Pacific Northwest was no less profound. As the Corps of Engineers and the Bureau of Reclamation built dam after dam on the Columbia and its tributaries — Grand Coulee, The Dalles, John Day, McNary, Chief Joseph, and dozens of others — BPA became the marketing agent for a hydroelectric system of staggering scale. At its peak, the Federal Columbia River Power System encompassed thirty-one federal dams with a combined generating capacity of over 22,000 megawatts, making it one of the largest hydroelectric systems in the world.

Like TVA, BPA was required to give preference in the sale of its power to public bodies and cooperatives. This preference power shaped the development of the Pacific Northwest's electricity sector in fundamental ways. Scores of public utility districts, municipal utilities, and rural electric cooperatives formed or expanded to purchase BPA power at cost-based rates that were, for decades, among the lowest

in the nation. The region's cheap electricity attracted energy-intensive industries — aluminum smelting, paper manufacturing, chemical production — that became pillars of the regional economy. The aluminum industry, in particular, was essentially a creation of BPA's low-cost hydropower; at one point, the Pacific Northwest produced roughly forty percent of the nation's primary aluminum.

BPA differs from TVA in important institutional respects. It does not own generating facilities; the dams are owned and operated by the Army Corps of Engineers and the Bureau of Reclamation, with BPA serving as the marketing and transmission agent. BPA does, however, own and operate one of the largest high-voltage transmission systems in the country — over 15,000 circuit miles of transmission lines that form the backbone of the Pacific Northwest's electric grid. BPA is organized as a federal agency within the U.S. Department of Energy, and like TVA, it is self-financing, covering its costs through power and transmission revenues rather than congressional appropriations. It is required by statute to set rates sufficient to recover its costs, including repayment to the U.S. Treasury of the federal investment in the dams and transmission system.

The Public-Versus-Private Power Debates

The creation of TVA, BPA, and the broader federal power program provoked one of the most sustained and bitter political controversies of the twentieth century. The private utility industry and its political allies denounced federal power as "creeping socialism," arguing that government had no business competing with private enterprise in the generation and sale of electricity. They contended that federal power projects enjoyed unfair advantages — tax exemptions, below-market borrowing costs, and freedom from state regulation — that made any rate comparison meaningless. They warned that the expansion of federal power would discourage private investment, undermine property rights, and set a dangerous precedent for government ownership of other industries.

Public power advocates responded with equal fervor. They argued that electricity was too important to be left to profit-driven monopolies, that the holding company abuses of the 1920s had demonstrated the dangers of private control, and that the refusal of private utilities to electrify rural America constituted a market failure of the first order. They pointed to the dramatic rate reductions and consumption increases in TVA's service territory as proof that private utilities had been overcharging consumers. They argued that the "yardstick" function of federal power was itself a public good, disciplining private rates even in areas not directly served by federal projects.

This debate, though it has waxed and waned over the decades, has never been fully resolved. It echoes through contemporary controversies over electricity restructuring, renewable energy policy, and the proper role of government in energy markets. The institutional legacy of the New Deal era — TVA, BPA, and the broader federal power establishment — remains embedded in the structure of the American grid, a permanent reminder that the market alone did not electrify America.

* * *

II. G&T Cooperatives: How Rural America Electrified

The Rural Electrification Act and the Birth of the Cooperative Model

If TVA and BPA addressed rural electrification through the supply side — building massive federal generating facilities and offering cheap wholesale power — the Rural Electrification Administration attacked the problem from the demand side, empowering rural communities to build their own distribution systems. The Rural Electrification Act of 1936 created the REA (housed initially as an executive agency, later transferred to the Department of Agriculture) and authorized it to make low-interest loans for the construction of electric distribution facilities in rural areas.

The genius of the REA program lay not in its technology but in its institutional innovation. The REA did not build or operate electric systems itself. Instead, it provided financing and technical assistance to locally organized rural electric cooperatives — nonprofit, member-owned entities formed by groups of farmers and rural residents who agreed to purchase electricity and pay for the construction of distribution lines serving their properties. Each cooperative was governed democratically, with each member holding one vote regardless of the amount of electricity consumed. Any surplus revenue beyond operating costs and debt service was returned to members as patronage capital — allocated to each member in proportion to their purchases and eventually retired (paid out in cash) according to the cooperative's financial capacity.

This cooperative model had deep roots in American agrarian tradition, drawing on the precedent of agricultural marketing cooperatives, mutual insurance companies, and other forms of collective self-help that had flourished in rural communities since the nineteenth century. But the application of the cooperative model to electricity distribution was new, and it required substantial support from the federal government to succeed. REA loans were offered at interest rates well below market — initially at the government's own cost of borrowing — and with repayment terms of up to thirty-five years. REA engineers developed standardized construction specifications that dramatically reduced the cost of rural line construction, and REA staff provided management training and organizational assistance to fledgling cooperatives.

The results were extraordinary. In 1935, only about eleven percent of American farms had electric service. By 1950, that figure had risen to roughly ninety percent. By 1960, rural electrification was essentially complete. The REA and the cooperatives it financed had accomplished in twenty-five years what private utilities had failed to achieve in half a century of exclusive franchise.

From Distribution to Generation and Transmission

The early rural electric cooperatives were purely distribution entities. They purchased wholesale power — often from the very private utilities that had declined to serve rural areas — and resold it to their

members. But this arrangement created a structural vulnerability: cooperatives were dependent on investor-owned utilities for their power supply, and those utilities could, within the bounds of wholesale rate regulation, charge cooperatives rates that limited the cooperatives' ability to keep retail rates low.

To address this vulnerability, distribution cooperatives in many regions banded together to form Generation and Transmission cooperatives — commonly known as G&T cooperatives or G&Ts. A G&T cooperative is owned by its member distribution cooperatives, just as a distribution cooperative is owned by its individual consumer-members. The G&T builds or acquires generation facilities, constructs transmission lines, and provides wholesale power to its member cooperatives under long-term contracts. This vertical integration gave rural electric cooperatives a measure of control over their power supply costs and insulated them from the pricing decisions of investor-owned wholesalers.

The formation of G&T cooperatives was facilitated by the same REA loan programs that had financed the original distribution systems. REA (and its successor, the Rural Utilities Service, or RUS, which absorbed REA's functions when the Department of Agriculture was reorganized in 1994) extended generation and transmission loans on favorable terms, enabling cooperatives to build coal-fired power plants, purchase shares of nuclear generating stations, and construct transmission infrastructure that connected their systems to the broader grid.

The All-Requirements Contract

The relationship between a G&T cooperative and its member distribution cooperatives is typically governed by an "all-requirements" contract — a long-term agreement under which the distribution cooperative commits to purchasing all of its wholesale power from the G&T. These contracts, which often extend for thirty-five to fifty years, provide the G&T with the revenue certainty needed to finance large capital investments in generation and transmission. The distribution cooperative, in return, receives the benefit of cost-based wholesale power and a voice in the governance of the G&T.

The all-requirements model has been a source of both strength and tension within the cooperative world. On the one hand, it provides stability and economies of scale, allowing relatively small rural systems to collectively own and operate large, efficient generating facilities. On the other hand, it can lock distribution cooperatives into purchasing power from a G&T even when cheaper alternatives may be available on the wholesale market — a concern that has grown increasingly acute as wholesale electricity markets have developed and as the cost of renewable energy has fallen. Some distribution cooperatives have sought to exit or renegotiate their all-requirements contracts, leading to disputes that have occasionally resulted in litigation and, in rare cases, the dissolution of the G&T relationship.

Major G&T Cooperatives and the Cooperative Landscape Today

The cooperative sector today encompasses approximately 830 distribution cooperatives and roughly sixty-three G&T cooperatives, collectively serving approximately forty-two million people across forty-seven states. Cooperatives account for roughly thirteen percent of the nation's electric distribution

and serve approximately fifty-six percent of the nation's landmass — a reflection of their origins in low-density rural areas.

Among the largest G&T cooperatives are entities of substantial scale and sophistication. Tri-State Generation and Transmission Association, headquartered in Westminster, Colorado, serves forty-two member cooperatives across four Western states and operates a diverse portfolio of coal, natural gas, and renewable generation. Basin Electric Power Cooperative, based in Bismarck, North Dakota, serves 141 member cooperatives across nine states and owns generation facilities ranging from lignite coal plants to wind farms. Associated Electric Cooperative in Springfield, Missouri, and Oglethorpe Power Corporation in Tucker, Georgia, are other major G&T systems with generating capacity measured in thousands of megawatts.

Challenges of the Modern Era

Rural electric cooperatives face a distinctive set of challenges in the contemporary electricity landscape. Their member demographics are changing: many cooperative service territories that were once purely agricultural now include exurban residential development, small cities, and commercial activity. Some cooperative territories have experienced significant population growth, while others — particularly in the Great Plains and rural South — have seen population decline and economic stagnation.

The cooperative business model, with its emphasis on cost-based pricing and member governance, creates both advantages and constraints. Cooperatives are generally well-regarded by their members for responsive service and community engagement, but they lack access to equity capital markets and must finance growth primarily through retained earnings (patronage capital) and debt. The RUS loan programs remain an important source of financing, but cooperatives increasingly access the private capital markets as well, issuing bonds through the National Rural Utilities Cooperative Finance Corporation (CFC) or commercial lenders.

The energy transition poses particular challenges for G&T cooperatives that invested heavily in coal-fired generation. Long-lived coal assets, financed with long-term debt, cannot be easily retired without stranding capital and imposing costs on member cooperatives. At the same time, the declining cost of wind and solar generation — resources that are often abundant in cooperative service territories — creates pressure from member cooperatives seeking access to cheaper, cleaner power. The tension between legacy generation commitments and the economics of new renewables has become one of the defining issues in the cooperative sector, testing the governance structures and contractual relationships that have held the cooperative system together for decades.

* * *

III. The Federal Footprint: How Federal Entities Interact with

Private Markets

The Power Marketing Administrations

The federal government's role in electricity generation and sales is administered primarily through four Power Marketing Administrations (PMAs), each responsible for marketing hydroelectric power from federal dams in a specific region of the country. These four agencies — the Bonneville Power Administration (BPA), the Western Area Power Administration (WAPA), the Southwestern Power Administration (SWPA), and the Southeastern Power Administration (SEPA) — together market power from over 130 federal hydroelectric projects with a combined installed capacity of approximately 55,000 megawatts.

BPA, as discussed above, markets power from the Federal Columbia River Power System in the Pacific Northwest. WAPA, headquartered in Lakewood, Colorado, markets power from fifty-seven federal hydroelectric plants in fifteen states across the central and western United States, including major Bureau of Reclamation projects such as Hoover Dam, Glen Canyon Dam, and the Central Valley Project in California. SWPA, based in Tulsa, Oklahoma, markets power from twenty-four Corps of Engineers dams in six states across the Southwest. SEPA, headquartered in Elberton, Georgia, markets power from twenty-two Corps of Engineers dams in eleven southeastern states.

Each PMA operates under the same fundamental statutory framework: federal hydroelectric power must be sold at rates that recover the full cost of production, including repayment of the federal investment, and preference in the sale of that power must be given to public bodies and cooperatives. The PMAs do not, with the partial exception of BPA, own or operate the generating facilities themselves; the dams are owned and operated by the Army Corps of Engineers or the Bureau of Reclamation, with the PMAs serving as the marketing and, in some cases, transmission agents.

The Army Corps of Engineers and the Bureau of Reclamation

The distinction between the entities that build and operate federal dams and the entities that market the power from those dams is a source of frequent confusion but is important to understanding the federal power system. The Army Corps of Engineers, a branch of the United States Army within the Department of Defense, is the nation's primary agency for water resources development in the eastern and central United States. The Corps builds and operates dams, locks, levees, and other water infrastructure for purposes including flood control, navigation, water supply, recreation, and hydroelectric power generation. Power generation is typically not the primary purpose of Corps projects but rather an incidental benefit of dams built for flood control or navigation — a distinction that has significant implications for cost allocation and power pricing.

The Bureau of Reclamation, an agency within the Department of the Interior, performs an analogous

function in the seventeen western states, where its primary mission is water supply and irrigation. The Bureau built many of the iconic dams of the American West — Hoover Dam, Grand Coulee Dam, Glen Canyon Dam, Shasta Dam — as components of large-scale water reclamation projects. Hydroelectric generation at these facilities is, again, ancillary to the primary irrigation and water supply mission, though the revenue from power sales has been critical to financing the Bureau's reclamation projects.

Together, the Corps and the Bureau own and operate federal hydroelectric facilities with a combined generating capacity of roughly 36,000 megawatts (exclusive of TVA), representing approximately half of the nation's total hydroelectric capacity and roughly three percent of total generation capacity. This federal hydroelectric fleet, though a relatively small share of total national generation, produces electricity at extremely low marginal cost — water, after all, is free — and thus represents some of the most economically valuable generation assets in the country.

Preference Power and the Wholesale Market

The preference power doctrine — the statutory requirement that federal hydroelectric power be sold first to public bodies and cooperatives — is the cornerstone of the federal power system and one of its most contentious features. The doctrine has its origins in the Reclamation Act of 1906 and has been codified in every major piece of federal power legislation since, including the Bonneville Project Act, the Flood Control Act of 1944, and numerous project-specific authorizations.

Under the preference doctrine, the PMAs must offer power from federal projects first to "preference customers" — publicly owned utilities, rural electric cooperatives, and certain other public entities — before making it available to investor-owned utilities or other non-preference customers. Preference customers receive power at cost-based rates determined by the PMAs, which are generally well below prevailing wholesale market prices. The difference between the cost-based preference rate and the market price represents a substantial economic benefit — in effect, a subsidy — that flows to the communities served by preference customers.

The economic value of federal preference power has grown enormously as wholesale electricity markets have developed and as the cost of alternative generation has risen relative to the zero-fuel-cost hydroelectric output of federal dams. In the Pacific Northwest, BPA's cost-based rates have historically been roughly half the wholesale market price, conferring an annual benefit to preference customers measured in billions of dollars. In the West, WAPA preference power from facilities like Hoover Dam and Glen Canyon Dam is similarly priced well below market, creating powerful incentives for existing preference customers to retain their allocations and for new public entities to seek preference status.

This arrangement is a source of persistent tension with investor-owned utilities and competitive power suppliers, who argue that the preference doctrine distorts wholesale markets by allocating low-cost power on the basis of institutional status rather than market competition. They contend that the below-market rates enjoyed by preference customers are effectively subsidized by federal taxpayers, who bear the risk of the federal investment in dams and receive a below-market return. They argue that the preference doctrine creates perverse incentives, discouraging preference customers from investing in

their own generation or pursuing energy efficiency because the opportunity cost of doing so — forgoing cheap federal power — is too high.

Defenders of the preference doctrine respond that federal hydroelectric projects were built with public funds for public purposes, and that it is entirely appropriate for the benefits of those investments to flow to publicly accountable entities rather than to private shareholders. They note that preference customers are required to pay cost-based rates that fully recover the federal investment, including interest, and that the "subsidy" alleged by critics is simply the difference between cost-based pricing and monopoly rents in imperfectly competitive wholesale markets. They argue that the preference doctrine has been a cornerstone of rural and small-community economic development for nearly a century and that its elimination would impose devastating rate increases on millions of Americans in areas that were historically underserved by private utilities.

Federal Power in an Era of Market Restructuring

The interaction between federal power entities and restructured wholesale markets has grown increasingly complex in recent decades. As regional transmission organizations (RTOs) and independent system operators (ISOs) have organized competitive wholesale markets across much of the country, the PMAs have been required to navigate the interface between their cost-based, preference-driven model and the market-based pricing mechanisms of organized wholesale markets.

BPA, for example, participates actively in the Western wholesale market, both as a seller of surplus power and as a buyer when its hydroelectric output is insufficient to meet its obligations to preference customers. WAPA similarly interacts with wholesale markets across the West and participates in the Western Energy Imbalance Market. These interactions raise difficult questions about how cost-based federal power should be integrated with market-based pricing, how the benefits of federal hydro should be allocated between preference customers and the broader market, and how federal entities should manage the revenue implications of market participation.

The federal power establishment also faces significant challenges related to the aging of its physical infrastructure. Many federal dams are now sixty to ninety years old, and their hydroelectric facilities require substantial investment in turbine rehabilitation, generator rewinding, and dam safety improvements. The costs of these investments must be recovered through the PMAs' cost-based rates, creating upward pressure on preference power prices even as the underlying resource — falling water — remains free. Climate change poses an additional challenge, as altered precipitation patterns and reduced snowpack threaten to diminish the hydroelectric output of federal dams in some regions, potentially reducing the volume of preference power available for sale and increasing its effective cost.

Despite these challenges, the federal power system remains a formidable presence in American electricity markets. The PMAs and TVA together market electricity from facilities with a combined capacity exceeding 80,000 megawatts, serve hundreds of wholesale customers, and operate transmission systems that are integral to the reliability of the national grid. Their preference power deliveries underpin the economics of hundreds of municipal utilities and rural electric cooperatives, supporting retail rates

that are, in many cases, among the lowest in the nation. The federal footprint in American electricity, established in the emergency of the Great Depression, has proved remarkably durable — a testament to the political and economic forces that created it and to the enduring appeal of cheap, publicly owned hydroelectric power.

* * *

Conclusion

The public power institutions examined in this chapter — TVA, BPA, the Power Marketing Administrations, and the rural electric cooperatives — represent a parallel tradition in American electricity, one that has coexisted with the investor-owned utility model for nearly a century. Born of the conviction that private enterprise alone could not or would not deliver universal electric service, these institutions brought light to rural America, powered the arsenal of democracy in World War II, and shaped the economic development of entire regions. They introduced into the American electricity system principles — preference power, cost-based pricing, cooperative ownership, democratic governance — that stand in deliberate contrast to the profit-driven model of investor-owned utilities.

Yet these institutions are not relics of a bygone era. They are active, consequential participants in contemporary electricity markets, controlling assets of enormous value, serving tens of millions of consumers, and grappling with the same challenges — decarbonization, grid modernization, market restructuring — that confront every participant in the American power system. The tensions that animated the public-versus-private power debates of the 1930s have not been resolved so much as transmuted into new forms: disputes over preference power allocations, arguments over cooperative contract flexibility, controversies over the proper role of federal entities in competitive markets. Understanding these tensions, and the institutions from which they arise, is indispensable to understanding the grid as it exists today — and as it will evolve in the decades to come.

* * *

Part III

Markets and Deregulation

Chapter 5: The Rise of the RTO and ISO

Introduction: The Great Unbundling

For most of the twentieth century, the American electric power system operated under a grand bargain. Vertically integrated utilities — companies that owned generating plants, transmission lines, and local distribution networks as a single enterprise — were granted exclusive franchises to serve defined territories. In exchange for this monopoly privilege, they submitted to rate regulation by state public utility commissions, which set prices based on the cost of service plus a reasonable rate of return on invested capital. The arrangement delivered remarkable results: near-universal electrification, improving reliability, and declining real costs through much of the postwar era. It was, by most measures, one of the most successful regulatory compacts in American economic history.

By the late 1980s, however, the consensus supporting this model had begun to fracture. The intellectual currents of deregulation, the practical experience of restructuring in other network industries, and mounting evidence that the vertically integrated model was producing inefficient outcomes combined to set the stage for a fundamental transformation. Over the span of roughly a decade — from the Energy Policy Act of 1992 through the full implementation of FERC Order 2000 in the early years of the new millennium — the federal government dismantled the core logic of the old system and replaced it with something radically different: a model in which competitive generators would sell power into organized wholesale markets, administered by independent entities that controlled the transmission grid without owning it.

This chapter tells the story of that transformation. It begins with the legal and intellectual origins of electricity restructuring and traces the two landmark regulatory orders — FERC Order 888 in 1996 and FERC Order 2000 in 1999 — that created the framework for competitive wholesale electricity markets. It then examines the institutional architecture of the Regional Transmission Organizations and Independent System Operators that emerged to operate these markets, exploring how they manage to coordinate one of the most complex physical systems ever built without owning any of its assets. Finally, it confronts the persistent challenge that arises whenever power must flow across the boundaries between these organizations — the so-called "seams problem" that remains one of the most vexing issues in

American electricity governance.

The transition from regulated monopoly to competitive markets was neither smooth nor complete. It produced spectacular failures alongside genuine efficiency gains. It created new institutions of enormous complexity and spawned stakeholder processes of sometimes stupefying procedural density. But it fundamentally reshaped the architecture of the American power system, and understanding its logic is essential to understanding how electricity markets function today.

* * *

5.1 FERC Order 888 and Order 2000: The Legal Big Bang

5.1.1 Intellectual and Political Origins of Restructuring

The movement to restructure the electricity industry did not emerge in a vacuum. It was the product of a broader intellectual revolution in the economics of regulation that gained force throughout the 1970s and 1980s. Economists of various political persuasions had grown increasingly skeptical of the traditional justification for rate-of-return regulation of natural monopolies. They argued that regulation itself created perverse incentives — encouraging overinvestment in capital (the so-called Averch-Johnson effect), suppressing innovation, and insulating incumbent firms from the discipline of competition. Where competitive markets could be made to work, they contended, the results would be superior to even the most well-intentioned regulatory oversight.

This intellectual current found its most dramatic practical expression in the deregulation of the airline industry in 1978 and the breakup of AT&T's telephone monopoly in 1984. Both episodes seemed to validate the economists' arguments. Airline deregulation produced lower fares and expanded service, at least in the aggregate. Telecommunications deregulation unleashed a wave of innovation and investment that would have been difficult to imagine under the old Bell System monopoly. These successes created a powerful political momentum: if competition could work in airlines and telephones, why not in electricity?

The analogy was seductive but imperfect. Electricity differs from airline seats and telephone calls in a fundamental physical respect: it cannot be economically stored at scale, it moves at the speed of light, and supply and demand must be balanced instantaneously across the entire network. These physical characteristics mean that the transmission grid — the high-voltage network that carries power from generators to load centers — exhibits genuine natural monopoly characteristics. Building duplicate transmission systems would be ruinously expensive and physically impractical. But the insight of the restructuring advocates was that generation — the actual production of electricity — was not a natural monopoly. Power plants could compete with one another, if they could gain non-discriminatory access to

the transmission network that connected them to customers.

This distinction between the naturally monopolistic transmission function and the potentially competitive generation function became the intellectual foundation of electricity restructuring. The goal was not to deregulate everything, but to unbundle the competitive segments of the industry from the monopoly segments and introduce market competition where it could work.

5.1.2 The Energy Policy Act of 1992

The first major legislative step toward restructuring came with the Energy Policy Act of 1992, known as EPAct 1992. This law, passed with bipartisan support, created a new category of power producer — the Exempt Wholesale Generator, or EWG — that could sell electricity at wholesale without being subject to the full panoply of regulations that applied to traditional utilities under the Public Utility Holding Company Act of 1935. More importantly, EPAct 1992 amended the Federal Power Act to give FERC the authority to order utilities to provide transmission service to third parties — a power known as "wheeling." This was a critical legal development. For the first time, the federal government had explicit statutory authority to force utilities to open their transmission lines to competitors.

The significance of this provision cannot be overstated. Under the old model, a vertically integrated utility that owned both power plants and transmission lines had an obvious incentive to deny or impede transmission access to competing generators. If an independent power producer built a more efficient plant next door, the incumbent utility could effectively neutralize that competitive threat by refusing to carry the rival's power across its wires, or by imposing unreasonable terms and conditions on transmission service. This practice — using transmission ownership as a competitive weapon — was the central problem that restructuring sought to address.

EPAct 1992 gave FERC the legal tools to attack this problem, but the Commission still needed to translate the statutory authority into specific rules. That task would consume the better part of four years and produce one of the most consequential regulatory proceedings in FERC's history.

5.1.3 FERC Order 888: Open Access and Functional Unbundling

On April 24, 1996, FERC issued Order No. 888, formally titled "Promoting Wholesale Competition Through Open Access Non-Discriminatory Transmission Services by Public Utilities." The order was the culmination of a massive rulemaking proceeding that generated tens of thousands of pages of comments from utilities, independent generators, state regulators, consumer advocates, and industrial customers. It represented FERC's most ambitious exercise of its authority under the Federal Power Act and fundamentally altered the structure of the wholesale electricity industry.

Order 888 rested on a factual finding that was itself a remarkable regulatory statement: FERC concluded that utilities had been engaging in widespread undue discrimination in the provision of transmission services. They had been using their control of the grid to favor their own generation over that of competitors, charging higher rates for third-party transmission, imposing more onerous

conditions, and providing inferior service quality. This finding of undue discrimination provided the legal predicate for the order's sweeping requirements.

The core mandate of Order 888 was the requirement that every public utility that owned, controlled, or operated transmission facilities file an Open Access Transmission Tariff, or OATT. The OATT was a standardized set of terms, conditions, and rates for transmission service that applied equally to the utility's own power transactions and to those of third parties. The principle was simple but revolutionary: a utility must offer transmission service to others on the same terms it provides to itself. No more favorable scheduling for the utility's own generators. No more creative interpretations of "available transmission capacity" that happened to exclude competitors. The transmission system was to become a common carrier, open to all on equal terms.

To enforce this principle, Order 888 required what FERC called "functional unbundling." Utilities were not required to divest their generation or transmission assets — that would have been corporate unbundling, a far more radical step that FERC did not believe it had the legal authority to mandate. Instead, they were required to separate their transmission operations from their power marketing functions, erecting internal walls to prevent the transmission side of the business from sharing non-public information with the generation and marketing side. The utility's merchants were to be treated as just another customer on the transmission system, with no informational advantages over competing generators.

Order 888 also addressed the thorny question of stranded costs — the investments that utilities had made under the old regulatory compact that might become uneconomic in a competitive market. Utilities that had built expensive nuclear plants or signed above-market purchased power contracts had done so with the expectation that they would recover those costs from captive customers. If those customers could now buy cheaper power from competitors using the utility's own transmission lines, the utility would be left with assets that could not earn a return — stranded costs. FERC concluded that utilities were entitled to recover legitimate and verifiable stranded costs, recognizing that a failure to do so would be both inequitable and legally problematic, as it would amount to a regulatory taking of property rights that had been created by prior regulatory commitments.

5.1.4 FERC Order 2000: The Regional Transmission Organization

Order 888 opened the transmission grid, but experience quickly revealed that functional unbundling within individual utilities was not sufficient to create truly competitive markets. Utilities found ways to comply with the letter of Order 888 while continuing to exercise subtle forms of discrimination. More fundamentally, the order left the transmission grid fragmented among dozens of individual utility operators, each managing its own piece of the network according to its own procedures. This fragmentation made it difficult for competitive generators and power marketers to transact across the broader regional grid, and it complicated the task of maintaining reliability across a physically interconnected system.

These concerns led FERC to issue Order No. 2000 on December 20, 1999. Where Order 888 had

focused on the terms of transmission service, Order 2000 focused on the institutional structure of grid management. The order defined a new organizational concept — the Regional Transmission Organization, or RTO — and strongly encouraged utilities to turn over operational control of their transmission facilities to these independent entities.

An RTO, as defined by Order 2000, was required to meet four minimum characteristics and perform eight minimum functions. The four characteristics were: independence from market participants (the RTO could not be owned or controlled by any entity with a financial interest in electricity transactions); appropriate scope and regional configuration (the RTO should be large enough to capture the benefits of regional coordination); operational authority over all transmission facilities under its control; and exclusive authority to maintain short-term reliability.

The eight minimum functions included: administering its own tariff and terms of service; managing congestion on the transmission system; developing and operating a market-based mechanism for congestion management (this would become the Locational Marginal Pricing system discussed in the next chapter); providing a transmission planning process; performing interconnection studies for new generators seeking to connect to the grid; ensuring interregional coordination; managing parallel path flows (the tendency of electricity to flow along all available paths, regardless of contractual arrangements); and maintaining an OASIS (Open Access Same-Time Information System) to provide real-time information about available transmission capacity.

Critically, Order 2000 encouraged but did not mandate RTO formation. FERC believed that it had the legal authority to require RTO participation, but it chose a more politically cautious approach, directing utilities to file proposals for RTO participation or, alternatively, to explain why they had chosen not to join an RTO. This voluntaristic approach reflected both FERC's reading of its legal authority and the political realities of the moment. Many utilities, particularly in the Southeast and West, fiercely opposed mandatory RTO participation, arguing that their existing bilateral trading arrangements worked well and that the costs of RTO formation would outweigh the benefits.

The result of this voluntary approach was a patchwork. In much of the Northeast, Mid-Atlantic, Midwest, and Texas, utilities did form or join RTOs. In the Southeast, much of the Mountain West, and parts of the Pacific Northwest, they did not. This geographic unevenness persists to the present day and represents one of the most significant structural features of the American power system.

5.1.5 FERC's Legal Authority and Its Limits

Both Order 888 and Order 2000 were exercises of FERC's authority under the Federal Power Act, originally enacted in 1935 and substantially amended by EPAct 1992. Understanding the scope and limits of this authority is essential to understanding the structure of electricity regulation in the United States.

FERC's jurisdiction under the Federal Power Act extends to the transmission and sale of electric energy at wholesale in interstate commerce. This jurisdictional hook — interstate commerce — is critically important. Because the alternating current grid is a synchronous machine in which power flows

cannot be confined to state boundaries, virtually all transmission is interstate in character. Even a transaction between a generator and a customer in the same state may involve power flows across the lines of utilities in neighboring states. This physical reality gives FERC an extraordinarily broad jurisdictional reach.

But the Federal Power Act also contains important limitations. Section 201(b) explicitly reserves to the states jurisdiction over facilities used in local distribution, retail sales to end-use customers, and the siting and construction of generation and transmission facilities. This means that FERC can regulate the wholesale price at which a generator sells power into the grid, but it cannot regulate the retail price at which a distribution utility sells that power to a household. It can set the rules for transmission service, but it cannot force a state to approve the construction of a new transmission line. It can create the framework for competitive wholesale markets, but it cannot compel states to restructure their retail markets.

This division of authority — sometimes called the "bright line" between federal and state jurisdiction, though in practice the line is anything but bright — has been a source of persistent tension. States that disagree with FERC's market-oriented approach have used their retained authority over generation siting, resource planning, and retail rate-setting to pursue alternative policies, sometimes in ways that conflict with FERC's wholesale market designs. The resulting jurisdictional complexity is a recurring theme throughout this book.

* * *

5.2 The "Traffic Cop" Model: Managing Power Without Owning It

5.2.1 The Organizational Architecture of an RTO/ISO

The Regional Transmission Organization is one of the most unusual institutional creations in American governance. It exercises enormous operational authority over a critical infrastructure system — dispatching power plants, managing billions of dollars in daily market transactions, and making real-time decisions that determine whether the lights stay on — yet it owns none of the physical assets it controls. It is not a government agency, though it is subject to extensive federal regulatory oversight. It is not a private corporation in the conventional sense, though it typically operates as a non-profit entity with a professional staff, a corporate headquarters, and an annual budget. It is, in essence, a creature of FERC — an organization that exists because the federal government decided that competitive electricity markets required an independent entity to perform functions that had previously been embedded within vertically integrated utilities.

The governance structure of an RTO reflects its peculiar institutional position. To meet FERC's

independence requirements, an RTO must be governed by a board of directors with no financial interest in any market participant. Board members cannot own stock in utilities, generators, or power marketers. They cannot have served as officers or employees of market participants within a specified period before their appointment. This independence requirement is designed to ensure that the RTO operates the grid and administers the markets in a neutral fashion, without favoring any particular class of market participant.

Beneath the board sits an elaborate stakeholder process through which market participants — utilities, generators, power marketers, industrial consumers, environmental organizations, and state regulators — provide input on the rules that govern the RTO's markets and operations. These stakeholder processes vary in their specific structures across the RTOs, but they typically involve multiple committees and working groups organized around functional areas such as market design, reliability, transmission planning, and resource adequacy. Proposals for rule changes generally originate in the stakeholder process, are debated and voted upon, and then filed with FERC for approval. FERC retains ultimate authority over all RTO tariff provisions and market rules, and it can reject stakeholder-approved proposals that it finds to be unjust, unreasonable, or unduly discriminatory.

A distinctive feature of RTO governance is the role of the Independent Market Monitor, or IMM. Each RTO is required to have an IMM — either an internal unit or an external contractor — whose function is to monitor market outcomes for evidence of market manipulation, the exercise of market power, or design flaws that produce inefficient results. The IMM analyzes bidding behavior, reviews market outcomes, and publishes regular reports on market performance. In some cases, the IMM has the authority to refer suspected market manipulation to FERC's Office of Enforcement for investigation and potential prosecution. The creation of the IMM function reflects a recognition that electricity markets, because of the physical characteristics of the grid and the inelasticity of short-term demand, are inherently susceptible to the exercise of market power — a topic explored in greater depth in subsequent chapters.

5.2.2 Core Functions of an RTO

The day-to-day operations of an RTO encompass a remarkable range of functions, from the split-second decisions of real-time dispatch to the multi-year horizon of transmission planning. These functions can be organized into several broad categories.

Reliability Coordination. The most fundamental responsibility of an RTO is maintaining the reliability of the bulk power system within its footprint. This means keeping the system in balance — ensuring that generation equals load plus losses at every instant — and maintaining the voltage and frequency of the alternating current within acceptable tolerances. The RTO's control room operators monitor the state of the grid continuously, using sophisticated energy management systems that process real-time data from thousands of sensors across the network. When contingencies occur — a generator trips offline, a transmission line is struck by lightning, demand surges unexpectedly — the operators must respond within seconds to rebalance the system and prevent cascading failures. This function is the

direct descendant of the utility control room operations that existed before restructuring, but it is now performed by an independent entity across a much larger geographic footprint.

Market Administration. RTOs operate organized wholesale electricity markets in which generators compete to sell power. The typical structure includes a day-ahead market, in which generators submit offers and loads submit bids for each hour of the following day, and a real-time market, which adjusts for deviations from the day-ahead schedule. The RTO uses a sophisticated optimization algorithm — typically a security-constrained economic dispatch with unit commitment — to determine which generators will run, at what output levels, and at what prices. The algorithm simultaneously considers generation costs, transmission constraints, and reliability requirements, producing a set of Locational Marginal Prices (LMPs) at each node on the network. These prices reflect the marginal cost of serving an additional megawatt of load at each location, including the costs of generation, transmission losses, and congestion. The mechanics of this pricing system are the subject of the next chapter; for present purposes, the key point is that the RTO administers a market of extraordinary computational complexity, clearing millions of dollars in transactions every hour.

In addition to energy markets, RTOs typically operate ancillary service markets for products such as regulation (the moment-to-moment balancing of supply and demand), spinning reserves (generators that are online and available to increase output within minutes), and non-spinning reserves (generators that can start up and deliver power within a specified time frame). Some RTOs also operate capacity markets — forward markets in which generators are paid for committing to be available during future periods of high demand. The design and performance of capacity markets is one of the most contentious topics in electricity policy, a subject addressed in later chapters.

Transmission Planning. RTOs are responsible for planning the expansion and reinforcement of the transmission system within their footprint. This is a complex undertaking that involves identifying future needs driven by load growth, generation retirements, new generator interconnections, and public policy requirements (such as state renewable energy mandates); evaluating alternative solutions; allocating costs among beneficiaries; and facilitating the construction of approved projects. Transmission planning in an RTO context is fundamentally different from the planning that occurred under the vertically integrated model. In the old model, a single utility planned its own system to meet its own customers' needs. In the RTO model, planning must account for the needs and interests of dozens or hundreds of market participants across multiple states, and cost allocation decisions inevitably create winners and losers, producing intense political conflict.

Interconnection Queue Management. New generators that wish to connect to the transmission system must apply to the RTO for an interconnection study, which determines the system upgrades needed to accommodate the new resource and allocates the costs of those upgrades. The interconnection queue — the pipeline of projects awaiting study and approval — has become one of the most significant bottlenecks in the American power system. As of the mid-2020s, the queues in most RTOs contain hundreds of gigawatts of proposed generation capacity, predominantly wind, solar, and battery storage projects, far exceeding any plausible level of actual development. The resulting delays — often spanning five years or more from application to commercial operation — have become a major impediment to the clean energy transition and have prompted FERC to undertake significant reforms of the interconnection

process.

Settlement. After the markets have cleared and power has flowed, the RTO must settle the financial transactions — determining how much each market participant owes or is owed based on their generation, consumption, and market positions. This settlement process involves massive data flows and complex calculations, and errors or disputes can take months to resolve.

5.2.3 Profiles of the Seven Major RTOs and ISOs

Seven major RTOs and ISOs currently operate in the United States, each with its own geographic footprint, membership composition, market design, and institutional culture. Understanding their distinguishing characteristics is important context for the discussions of market design and policy that follow.

PJM Interconnection is the largest RTO in terms of both geographic scope and market volume. Its footprint stretches from New Jersey westward through Pennsylvania, Ohio, and Indiana, and southward into Virginia, West Virginia, and parts of the Carolinas and the District of Columbia. PJM serves approximately 65 million people across thirteen states and administers the largest competitive wholesale electricity market in the world, with annual billings exceeding forty billion dollars. PJM traces its origins to a power pool formed in 1927 and became an ISO in 1997, making it one of the earliest adopters of the RTO model. It operates robust day-ahead and real-time energy markets, ancillary service markets, and a capacity market known as the Reliability Pricing Model, or RPM, which has been a subject of frequent regulatory controversy.

The Midcontinent Independent System Operator (MISO) covers a vast geographic footprint stretching from Manitoba and the upper Midwest southward through Indiana, portions of the Great Plains, and into Mississippi, Louisiana, Arkansas, and Texas. This north-south corridor makes MISO distinctive in its geographic configuration and creates unique challenges for transmission planning and resource adequacy, as the northern and southern portions of the footprint have different load patterns, resource mixes, and state regulatory environments. MISO operates energy and ancillary service markets and a resource adequacy construct, though its capacity market design differs significantly from PJM's.

The Southwest Power Pool (SPP) covers a large portion of the central United States, from the Dakotas and Nebraska southward through Kansas, Oklahoma, and into portions of the Texas Panhandle, Arkansas, and Louisiana. SPP's footprint is characterized by exceptional wind resources, and the organization has at times seen wind energy provide more than seventy percent of the electricity on its system. SPP was originally a reliability coordinator and evolved into a full RTO with organized energy markets, implementing its Integrated Marketplace in 2014.

The Electric Reliability Council of Texas (ERCOT) is unique among the major grid organizations in two important respects. First, the Texas grid is not synchronously interconnected with the Eastern or Western Interconnections — it operates as an electrical island connected to the rest of the country only through limited direct current ties. This physical separation has a profound legal consequence: because power does not flow in interstate commerce across the ERCOT system, FERC does not have jurisdiction

over ERCOT's wholesale markets. ERCOT is regulated by the Public Utility Commission of Texas rather than FERC. Second, ERCOT operates an energy-only market without a capacity market, relying on scarcity pricing — the prospect of very high energy prices during periods of tight supply — to incentivize investment in new generation. This design choice was tested dramatically during Winter Storm Uri in February 2021, when extreme cold caused widespread generator outages and produced catastrophic system failures that left millions of Texans without power for days.

The New York Independent System Operator (NYISO) administers wholesale electricity markets for New York State. The New York system is characterized by persistent transmission congestion between upstate, where hydroelectric and nuclear plants provide relatively low-cost energy, and downstate, where the densely populated New York City metropolitan area drives high demand but faces constraints on local generation. NYISO operates energy, ancillary service, and capacity markets and has been at the forefront of efforts to integrate state clean energy policies into wholesale market design.

ISO New England (ISO-NE) covers the six New England states: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. The New England system faces distinctive challenges related to its dependence on natural gas for electricity generation and the constraints on natural gas pipeline capacity, particularly during winter heating season when residential gas demand competes with electric generation for pipeline space. ISO-NE operates a Forward Capacity Market that has undergone multiple design revisions in response to state policies promoting renewable energy and the retirement of older fossil-fuel and nuclear generators.

The California Independent System Operator (CAISO) manages the grid for most of California and has extended its real-time market to cover portions of neighboring western states through the Western Energy Imbalance Market (WEIM), a real-time trading platform launched in 2014 that allows participating utilities in the West to optimize dispatch across a broader geographic area without joining a full RTO. CAISO's footprint is distinguished by the highest penetration of solar generation in the country, which creates the now-famous "duck curve" — a dramatic midday depression in net load followed by a steep evening ramp as solar output declines and demand increases. Managing this pattern has made CAISO a laboratory for the operational challenges of integrating variable renewable energy at scale.

* * *

5.3 Retail Restructuring and Customer Choice

5.3.1 The Extension of Competition to Retail Markets

The restructuring story told thus far has focused on wholesale markets — the institutional framework

through which generators sell power to load-serving entities. But in many states, restructuring extended beyond wholesale markets to the retail level, giving end-use customers the ability to choose their electricity supplier in much the same way they might choose a long-distance telephone carrier or a natural gas marketer. Understanding this dimension of restructuring is important both because it directly affects the prices paid by tens of millions of American households and businesses and because the political experience of retail restructuring — its successes, its failures, and its scandals — has profoundly shaped the trajectory of electricity policy.

5.3.2 The Logic of Retail Choice

The case for retail choice follows naturally from the case for wholesale competition. If the generation of electricity is not a natural monopoly — if multiple generators can compete to supply power — then why should customers be forced to buy from a single distribution utility at a regulated price? Why not allow competitive retail electric providers (sometimes called alternative retail suppliers, or competitive energy suppliers) to purchase power on the wholesale market and resell it to customers at prices and on terms of their own design?

In a retail choice market, the local distribution utility retains its monopoly over the physical delivery of electricity — the wires, transformers, and meters that carry power from the substation to the customer's premises. The distribution utility remains regulated by the state public utility commission and continues to charge a regulated delivery rate. But the commodity itself — the electrons — can be procured from a competitive supplier chosen by the customer. The customer's bill is split into two components: a regulated distribution charge paid to the local utility and a supply charge paid to the chosen competitive supplier (or to the utility under a "default service" or "provider of last resort" arrangement for customers who do not affirmatively choose a competitive supplier).

The theoretical benefits of retail choice mirror those claimed for competition generally: downward pressure on prices through supplier competition, product innovation (such as fixed-price contracts, green energy offerings, time-of-use plans, and bundled energy management services), and the discipline of customer choice forcing suppliers to be more responsive to customer preferences.

5.3.3 The Geography of Retail Choice

Retail choice is not available everywhere. It exists only in states that have enacted enabling legislation, and the map of retail choice states reflects the political geography of electricity restructuring. The states that adopted retail choice were concentrated in the Northeast, Mid-Atlantic, and parts of the Midwest and Texas — regions that were generally served by higher-cost utilities and that were more receptive, for various political and economic reasons, to the restructuring agenda of the 1990s.

Texas adopted retail choice as part of its comprehensive restructuring legislation, Senate Bill 7, in 1999. Within the ERCOT footprint, residential and commercial customers in areas served by investor-owned utilities can choose from dozens of competitive retail electric providers offering a variety

of plans, price structures, and contract terms. The Texas retail market is widely considered the most active and competitive in the nation, with high customer switching rates and a vibrant marketplace of product offerings.

In the Northeast, retail choice was adopted by most states in the region, including Pennsylvania, New York, Connecticut, Massachusetts, New Jersey, Maryland, and Ohio. The experience in these states has been mixed. In some markets, competitive suppliers have offered savings relative to the default utility rate, particularly for large commercial and industrial customers with the sophistication to evaluate competing offers and the load profiles that make them attractive to competitive suppliers. For residential customers, the results have been less uniformly positive. Several states have experienced controversies involving competitive suppliers whose introductory "teaser" rates gave way to prices significantly above the default service rate, and consumer protection concerns have prompted regulatory investigations and enforcement actions.

The Southeast, much of the Mountain West, and the Pacific Northwest have not adopted retail choice. In these regions, customers continue to purchase electricity from their local utility — whether investor-owned, municipal, or cooperative — at rates set by state regulators or utility governing boards. The persistence of the traditional model in these regions reflects a combination of relatively low electricity prices (which reduced the political impetus for restructuring), the political influence of vertically integrated utilities, and skepticism about whether the benefits of competition demonstrated in other industries would materialize in electricity.

5.3.4 The California Crisis and Its Shadow

No account of retail restructuring is complete without acknowledging the event that nearly derailed the entire enterprise: the California electricity crisis of 2000–2001. California was among the first states to implement retail choice, under legislation enacted in 1996. The restructuring was accompanied by a mandatory rate reduction for residential customers and a freeze on utility rates that was intended to allow utilities to recover their stranded costs.

The crisis that erupted in the summer of 2000 resulted from a confluence of factors: a flawed market design that gave generators market power, drought conditions that reduced hydroelectric output, tight natural gas supplies, emission constraints that limited the operation of older plants, and the manipulation of wholesale markets by traders at Enron and other energy companies. Wholesale prices spiked to extraordinary levels, but the retail rate freeze prevented utilities from passing these costs through to customers. Pacific Gas and Electric Company was driven into bankruptcy. Rolling blackouts swept the state. The political fallout was immense — Governor Gray Davis was recalled from office in part because of the crisis, and California effectively abandoned retail competition, suspending direct access for a decade.

The California crisis cast a long shadow over the retail choice movement nationwide. Several states that had been considering or implementing restructuring paused or reversed course. The crisis was frequently invoked by opponents of retail competition as evidence that electricity markets were

fundamentally unsuited to competitive provision — a characterization that proponents argued was unfair, given the specific design flaws and market manipulation that had produced the California outcome. But the political damage was done. The momentum toward universal retail choice, which had seemed almost irresistible in the late 1990s, was broken.

5.3.5 The Current Landscape

Today, retail choice exists in approximately fifteen to twenty states, depending on how partial or limited-access programs are counted. In most of these states, the competitive retail market is well-established and functioning, though the degree of customer participation varies widely. Large commercial and industrial customers are far more likely to have chosen a competitive supplier than residential customers, reflecting both the greater economic incentive (larger electricity bills provide more savings potential) and the greater ability to evaluate complex supply offers.

The retail choice landscape has evolved in recent years as competitive suppliers have expanded their product offerings beyond simple commodity pricing. Green energy plans, in which the supplier matches the customer's consumption with renewable energy certificates, have become a significant and growing market segment. Fixed-price contracts that insulate customers from wholesale price volatility have proven popular. Some competitive suppliers offer bundled services that include home energy management, smart thermostat installation, and demand response participation.

The coexistence of retail choice states and traditionally regulated states within the same RTO footprint creates an additional layer of complexity in electricity governance. In PJM, for example, some states within the footprint have active retail choice markets (Pennsylvania, Ohio, New Jersey, Maryland) while others retain traditional regulation (Virginia, portions of the Carolinas). The wholesale market administered by PJM operates identically across these states, but the retail experience of customers differs fundamentally depending on which side of a state line they happen to live on.

5.3.6 The Non-RTO Landscape: How the Other Half Operates

The map of RTO and ISO territories reveals a striking feature that is easy to overlook from within the restructured world: a vast portion of the American grid operates entirely outside the RTO framework. Approximately forty percent of U.S. electricity load is served by utilities that have not joined an RTO or ISO — a share that encompasses much of the Southeast, significant portions of the Mountain West, and parts of the Pacific Northwest. Understanding how these regions manage their power systems is essential to any complete picture of the American grid, because the non-RTO landscape is not a relic of an older era awaiting modernization. It is, in the view of its participants, a deliberate and defensible alternative to the organized market model.

The defining institution of the non-RTO landscape is the vertically integrated electric utility. In states such as Georgia, Alabama, Florida, South Carolina, and much of the Carolinas, utilities like Southern Company, Duke Energy, and Dominion Energy continue to own generation, transmission, and

distribution as a unified enterprise. They do not participate in centralized wholesale markets for their day-to-day operations. Instead, they generate most of the power their customers consume from their own fleet of power plants, supplemented by bilateral contracts with neighboring utilities and independent generators negotiated through direct, private agreements rather than through a competitive auction.

This bilateral trading model differs fundamentally from the RTO approach. Where an RTO clears supply and demand through a centralized algorithm that produces transparent Locational Marginal Prices at thousands of nodes, the bilateral model relies on individually negotiated contracts between willing buyers and sellers. The terms — price, volume, delivery schedule, risk allocation — are set through direct negotiation, and the resulting prices are generally not publicly disclosed. Proponents of the bilateral model argue that it provides price stability and planning certainty that short-term organized markets cannot match, because utilities can lock in fuel supplies and generation costs years in advance. Critics counter that the absence of transparent price signals conceals inefficiencies and shields incumbent generators from competition that would drive down costs.

Reliability coordination in the non-RTO world takes a different institutional form as well. Rather than the centralized dispatch and reliability management provided by an RTO, non-RTO utilities rely on a combination of reserve sharing agreements, joint planning committees, and third-party reliability coordinators. The Southeast, for example, is covered by several reliability coordination arrangements, including the Southeast Reliability Corporation (a NERC regional entity) and various reserve sharing groups through which utilities agree to provide emergency assistance to one another. These arrangements have generally maintained reliable service — the Southeast does not suffer from systematically higher outage rates than RTO regions — though some analysts argue that the lack of transparent market signals makes it harder to assess whether the region's reliability comes at an efficiently low cost or whether customers are paying more than necessary for the security they receive.

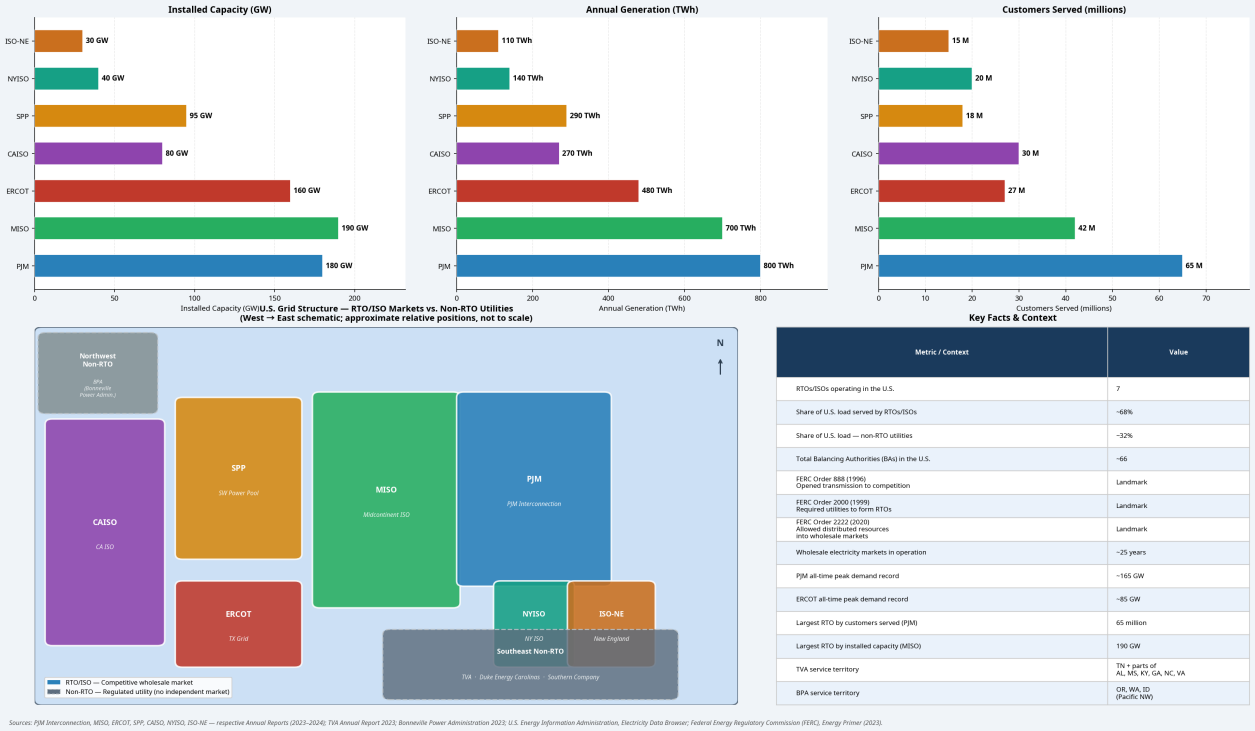

Figure 5.1: U.S. Grid Structure — RTO/ISO Markets vs. Non-RTO Utilities (Source: FERC, EIA, RTO Annual Reports)

The political economy of RTO non-participation is complex and deeply entrenched. Large vertically integrated utilities wield considerable political influence in their home states, where they are often the largest employer, the largest taxpayer, and the most generous campaign contributor. State public utility commissions in the Southeast have historically been sympathetic to utility arguments that joining an RTO would surrender local control over energy policy to a distant, federally regulated bureaucracy. The argument carries weight: in an RTO, transmission planning and wholesale market rules are set through a stakeholder process governed by FERC, not by state regulators. For states accustomed to exercising direct authority over their utilities' investment decisions, the prospect of ceding that authority to a regional body — dominated, perhaps, by interests from other states — is unappealing regardless of the potential economic benefits.

The result is a persistent geographic divide in American electricity governance. Roughly speaking, the northern, midwestern, and western portions of the country operate under the organized market model, while the Southeast and scattered portions of the West operate under the traditional bilateral model. The boundary between these two worlds is not merely an administrative curiosity. It has real consequences for consumers, generators, and the trajectory of the energy transition. Renewable energy developers, for instance, report that it is significantly more difficult to interconnect new wind and solar projects in non-RTO regions, where the interconnection process is managed by individual utilities rather than by a centralized queue with transparent procedures. And the absence of organized capacity markets in non-RTO regions means that the economic signals guiding investment in new generation are shaped primarily by utility integrated resource plans and state regulatory proceedings rather than by competitive

market outcomes.

Whether the non-RTO model will persist indefinitely is an open question. FERC's authority to mandate RTO participation remains limited by the political constraints that prevented Order 2000 from requiring membership. Periodic proposals to extend organized markets to the Southeast — including recent discussions about an energy imbalance market modeled on CAISO's Western EIM — have generated interest but have not yet produced fundamental change. The utilities that operate outside RTOs are not standing still; many have adopted elements of the organized market model, such as real-time balancing markets and transparent interconnection procedures, without submitting to full RTO membership. The American grid may ultimately converge on a common market structure, or it may continue to operate as a patchwork of organized and bilateral systems for decades to come. In either case, the coexistence of these two models — and the tensions between them — is one of the defining features of the grid as it exists today.

5.3.7 Purchased Power and Economy Energy: Trading Across the Boundary

The boundary between the RTO and non-RTO worlds is not a wall. It is a seam — porous, busy, and governed by a web of contracts, tariffs, and operating agreements through which enormous quantities of electricity flow every day. Understanding how and why power is bought and sold across this boundary illuminates the economic logic that connects the two halves of the American grid.

The fundamental concept is simple. Every utility, whether it operates within an RTO or outside one, maintains a dispatch stack — a ranked list of available generators ordered by their marginal cost of production. When demand rises, the utility dispatches the next cheapest available generator. But what if the next generator in the utility's own stack costs $65 per megawatt-hour, while the price at the interface with a neighboring system — whether an RTO's transparent LMP or a bilateral offer from an adjacent utility — is $40 per megawatt-hour? The arithmetic is obvious: buying power from the neighbor saves $25 per megawatt-hour, a savings that flows directly to ratepayers as a reduction in fuel and purchased power expense.

This is the concept of **economy energy** — power purchased not because the buying utility lacks sufficient generation capacity to serve its load, but because external power is cheaper than the next unit the utility would otherwise dispatch. Economy energy transactions are distinct from emergency power purchases (which occur when a utility faces a genuine capacity shortfall) and from long-term power purchase agreements (which lock in supply over years or decades). Economy energy is opportunistic, short-term, and driven entirely by the real-time comparison of internal marginal cost against the external market price.

Consider the Tennessee Valley Authority, the nation's largest public power provider, serving approximately ten million people across seven states. TVA is not a member of any RTO. It operates its own diverse fleet of nuclear, natural gas, coal, hydroelectric, and solar generation, dispatching these resources to serve the load of its 153 local power companies. TVA's service territory borders PJM to the north and east, and MISO to the west — two of the largest organized wholesale markets in the world.

When TVA's next available generator is a natural gas combustion turbine with a heat rate and fuel cost that produces a marginal cost of $60 per megawatt-hour, and PJM's LMP at the interface is $35 per megawatt-hour, TVA can purchase economy energy from PJM at the interface price, avoid starting its own expensive peaking unit, and pass the savings through to its ratepayers. The transaction is settled through the interchange scheduling process described in Chapter 2 — TVA schedules an import, adjusts its ACE calculation accordingly, and the power flows across the interconnection ties that link TVA's transmission system with PJM's.

The reverse also occurs. When TVA has surplus low-cost generation — perhaps its nuclear fleet is producing at full output during a period of low demand — it can sell economy energy to PJM or MISO if the market price exceeds TVA's marginal cost. In this case, TVA's ratepayers benefit from the revenue earned on the sale, which offsets the cost of the generation that would have been curtailed or idled.

These transactions happen routinely across every major interface on the grid. Within RTOs, the optimization is automatic — the SCED algorithm dispatches the cheapest available generation across the entire RTO footprint, effectively making every intra-RTO transaction an economy energy transaction optimized by software. Between RTOs and non-RTO utilities, the process is more manual: system operators on both sides communicate available capacity and prices, schedule interchanges in advance (typically on an hourly basis through the OASIS electronic bulletin board system), and settle the transactions after the fact.

The volume of purchased power in the American electricity system is enormous. According to EIA data, purchased power expense — electricity bought from external sources rather than generated from the utility's own plants — typically represents 20 to 40 percent of total power supply costs for large vertically integrated utilities. For some smaller utilities and cooperatives that own little or no generation of their own, purchased power is the dominant cost category. Even for a generation-rich system like TVA, economy energy purchases and sales at the margins of the dispatch stack can save ratepayers hundreds of millions of dollars annually.

The existence of these cross-boundary transactions has important implications for the debate over grid structure. Proponents of RTO expansion argue that economy energy transactions between RTOs and non-RTO utilities demonstrate the value of transparent price signals and coordinated dispatch — and that formalizing these arrangements through RTO membership would capture even greater savings by optimizing dispatch continuously rather than through periodic bilateral scheduling. The non-RTO utilities counter that they already capture the available savings through their existing bilateral arrangements, without surrendering operational control to a federally regulated market administrator.

The truth lies somewhere between these positions. Bilateral economy energy transactions do capture some of the gains from trade, but they are inherently less efficient than the continuous, algorithm-driven optimization that occurs within an RTO. Bilateral scheduling typically occurs on an hourly basis, while RTO dispatch is optimized every five minutes. Bilateral transactions require human operators to identify opportunities, negotiate terms, and schedule interchanges — a process that is slower and less granular than algorithmic optimization. Studies commissioned by various RTOs have estimated that the efficiency gains from moving to centralized dispatch from bilateral arrangements range from $1 to $3 per megawatt-hour served — seemingly small numbers that, applied to the hundreds of terawatt-hours

consumed in non-RTO territories, aggregate to billions of dollars annually.

* * *

5.4 The Seams Problem: Friction at the Borders

5.4.1 The Nature of the Problem

The map of American RTOs and ISOs reveals an inconvenient truth: the boundaries between these organizations are drawn along institutional and political lines that bear no necessary relationship to the physical flow of electricity. Electrons do not respect corporate jurisdictions. Power generated in PJM may flow through MISO territory to reach a customer in SPP. A transaction between a generator in NYISO and a load in ISO-NE may be physically supported by power flows across PJM's network. The laws of physics — specifically, Kirchhoff's laws governing current flow in electrical networks — ensure that electricity takes all available paths from source to sink, regardless of which RTO or utility happens to own the wires along each path.

This physical reality creates what industry participants and regulators have come to call the "seams problem." The term refers to the friction, inefficiency, and coordination challenges that arise at the boundaries — the seams — between adjacent RTOs. These challenges take several forms.

Different Market Rules and Prices. Each RTO operates its own markets with its own rules, its own software, and its own price formation mechanisms. Even though the underlying principles of Locational Marginal Pricing are similar across RTOs, the specific implementation details — how constraints are modeled, how reserves are procured, how uplift costs are allocated — can differ significantly. The result is that the price of electricity at a node just inside the PJM border may be substantially different from the price at a node just across the border in MISO, even though the two nodes are physically adjacent on the same transmission line. These price differences may reflect genuine differences in supply and demand conditions, but they may also reflect modeling inconsistencies, differences in market timing, or coordination failures between the two RTOs.

Pancaked Transmission Rates. When a wholesale transaction spans two RTOs, the parties may be required to pay separate transmission charges to each RTO — a practice known as rate pancaking. A generator in MISO selling power to a customer in PJM must pay for transmission service on MISO's system and again on PJM's system. This double charge can make inter-regional transactions uneconomic even when they would be efficient from a system-wide perspective. The pancaking problem is a direct consequence of the institutional fragmentation of the grid: each RTO recovers its transmission costs through its own tariff, and transactions that cross RTO boundaries must pay into both tariff structures. FERC has taken steps to eliminate rate pancaking in some contexts — for example, by requiring the use

of a single transmission rate for transactions within an RTO regardless of how many utility transmission zones the power crosses — but the inter-RTO pancaking problem has proven more resistant to solution.

Coordination of Real-Time Operations. Maintaining reliability across the seam between two RTOs requires continuous coordination between their respective control rooms. When a contingency occurs near the border — a generator trips or a transmission line goes out of service — both RTOs may be affected, and their responses must be coordinated to avoid conflicting actions that could make the situation worse. This coordination is complicated by the fact that each RTO may be using different models of the system, different assumptions about operating limits, and different procedures for responding to contingencies.

Transmission Planning Across Seams. Perhaps the most consequential aspect of the seams problem is the difficulty of planning and building transmission infrastructure that spans multiple RTOs. Each RTO has its own transmission planning process, its own criteria for identifying needed projects, and its own cost allocation methodology. A transmission line that would provide significant benefits to customers in both MISO and PJM requires approval and cost allocation decisions from both organizations — a process that is bureaucratically complex, politically contentious, and painfully slow. The result is a systematic underinvestment in inter-regional transmission, a deficiency that multiple studies have identified as a significant drag on economic efficiency and a barrier to the integration of renewable energy resources.

5.4.2 Joint Operating Agreements and Seams Coordination

To manage the challenges at their borders, adjacent RTOs enter into Joint Operating Agreements, or JOAs. These agreements establish protocols for the coordination of real-time operations, the exchange of data and market information, the management of parallel flows, and the resolution of disputes. The JOA framework represents an attempt to achieve some of the benefits of coordinated operations across RTO boundaries without requiring the much more difficult step of full organizational integration.

The most developed JOA relationships are typically between RTOs that share long and electrically significant boundaries. The PJM-MISO JOA, for example, covers coordination along a border that stretches from Michigan to Virginia and across which enormous volumes of power flow daily. The agreement establishes procedures for coordinated transaction scheduling, congestion management at the seam, and the allocation of costs associated with managing parallel flows. Similarly, the PJM-NYISO JOA addresses the heavily congested interfaces between PJM's system and the New York Control Area.

JOAs have achieved genuine improvements in seams coordination, but they remain limited instruments. They are, fundamentally, contracts between independent organizations, each of which retains its own governance, its own market rules, and its own institutional incentives. The negotiations that produce JOAs are complex and often contentious, as each RTO seeks to protect the interests of its own stakeholders. Amendments to JOAs must typically be approved by both RTOs' stakeholder processes and filed with FERC, a procedure that can take years to complete.

One specific coordination mechanism that has shown promise is the use of coordinated transaction

scheduling at RTO seams. Under this approach, the two adjacent RTOs share price information before clearing their respective markets, allowing traders to submit interface bids that reflect the expected price difference between the two systems. This helps to ensure that power flows across the seam in the economically efficient direction — from the lower-priced system to the higher-priced system. More recently, some RTOs have explored or implemented "tie optimization" protocols that go further, allowing the RTO dispatch algorithms to directly account for conditions on the other side of the seam when determining optimal flows across the interface.

5.4.3 Why Full Integration Remains Elusive

Given the costs and inefficiencies created by seams, a natural question arises: why not simply merge adjacent RTOs into larger organizations, eliminating the seams entirely? The question is reasonable, and various proposals for RTO consolidation have been advanced over the years. FERC itself has periodically suggested that a smaller number of larger RTOs might be more efficient than the current patchwork.

But full integration has proven extraordinarily difficult to achieve for several reasons. First, the existing RTOs have developed distinct market designs, software systems, and operating procedures over many years, and harmonizing these systems would be a massive and expensive undertaking. The costs and risks of integration — including the potential for software failures, market disruptions, and operational errors during the transition — are concrete and immediate, while the benefits of integration, though potentially large, are diffuse and uncertain.

Second, RTO consolidation raises difficult governance questions. Each existing RTO has a stakeholder community that has invested years in developing market rules tailored to its region's specific characteristics. Generators in MISO's wind-rich footprint have different interests than generators in PJM's more diverse resource mix. State regulators in New England face different challenges than those in the Southwest. Merging two RTOs means merging two stakeholder communities, with all the political complexity that entails. Decisions about market design, cost allocation, and reliability standards that were previously made within a relatively homogeneous stakeholder group must now be negotiated among a larger and more diverse set of interests.

Third, and perhaps most fundamentally, RTO boundaries reflect underlying political realities. The regions that have not joined RTOs — particularly the Southeast — have resisted doing so for reasons that go beyond technical or economic analysis. Utilities in these regions have well-established relationships with their state regulators, and both parties may prefer the existing bilateral model to the perceived uncertainties and loss of local control that come with RTO membership. FERC's decision in Order 2000 not to mandate RTO participation reflected a judgment that forcing these regions into RTOs would provoke a political backlash that could undermine the entire restructuring enterprise.

The result is that the seams problem remains a persistent feature of the American power system — a tax on efficiency that market participants, regulators, and RTOs themselves continue to work to reduce but have not been able to eliminate. The most promising developments in recent years have come not

from organizational consolidation but from the expansion of limited coordination mechanisms — like CAISO's Western Energy Imbalance Market, now known as the Western Energy Imbalance Service — that allow utilities outside of full RTOs to participate in real-time dispatch optimization. These arrangements represent an intermediate step between the fully bilateral world and full RTO membership, offering some of the benefits of coordinated dispatch without requiring the full institutional commitment of RTO participation. Whether they represent a waystation on the road to eventual full integration or a durable institutional equilibrium remains an open question.

* * *

Conclusion: An Unfinished Revolution

The creation of RTOs and ISOs represents one of the most ambitious experiments in the governance of critical infrastructure ever undertaken in the United States. In the span of roughly a decade, the federal government transformed the wholesale electricity industry from a system of regulated bilateral transactions among vertically integrated utilities into a system of organized competitive markets administered by independent entities with no ownership stake in the assets they control. The intellectual audacity of this transformation — asking competitive markets to coordinate a physical system that must be balanced in real time, across thousands of miles of transmission lines and thousands of generating units — is easy to underestimate from the vantage point of the present, when the RTO model has become the familiar backdrop of electricity policy debates.

The results have been mixed. In the regions where RTOs operate, competitive wholesale markets have generally produced efficiency gains through improved dispatch of generating resources, better utilization of the transmission system, and more transparent price signals. The RTO model has also facilitated the integration of renewable energy resources by providing a framework for managing the variability and uncertainty of wind and solar generation across a large geographic area. At the same time, the RTO model has introduced new forms of complexity, new opportunities for market manipulation, and new regulatory challenges that would have been unimaginable under the old vertically integrated model.

The revolution, moreover, is incomplete. Roughly a third of the country's electricity load remains outside of organized RTO markets. The seams between existing RTOs continue to impede efficient power flows and complicate transmission planning. The interaction between FERC's wholesale market authority and state retail and resource planning authority generates persistent jurisdictional friction. And the fundamental question of whether competitive markets can deliver the long-term investment in generation and transmission infrastructure needed to maintain reliability and achieve clean energy goals remains a subject of vigorous and unresolved debate.

The chapters that follow will explore these questions in detail, examining the mechanics of

electricity pricing, the design of capacity markets, and the governance challenges that arise when competitive markets must coexist with public policy mandates. But the institutional foundation — the RTO as traffic cop, market administrator, and reliability coordinator — is now in place. Understanding how it was built, how it works, and where its limits lie is essential to understanding everything that follows.

* * *

Chapter 6: Energy Markets vs. Capacity Markets

Introduction: The Impossible Commodity

Electricity is unlike any other commodity traded in modern markets. It cannot be economically stored at scale. It travels at the speed of light across a shared network whose physics no single party controls. It must be produced and consumed in the same instant, with supply and demand balanced not roughly, not approximately, but exactly — continuously, every second of every day — or the entire system collapses. And yet, beginning in the 1990s, the United States undertook one of the most ambitious experiments in economic policy: it attempted to create competitive markets for this impossible commodity, replacing the administered pricing of vertically integrated monopolies with the decentralized decision-making of buyers, sellers, and the price mechanism.

The results have been remarkable, frustrating, and deeply contested. Wholesale electricity markets have driven efficiency gains, reduced fuel costs, and revealed through price signals what decades of utility planning often obscured — the true marginal cost of delivering a megawatt-hour of energy to a specific location at a specific moment in time. But these same markets have also exposed a set of economic puzzles that remain unresolved after more than two decades of operation. Chief among them is a deceptively simple question: Can a market for electricity, even a well-designed one, produce prices that are simultaneously efficient in the short run and sufficient in the long run to ensure that the lights stay on?

This chapter examines the two interlocking market structures that have emerged from this question. The first is the energy market, where generators compete to supply electricity in real time and prices are set through a mechanism known as Locational Marginal Pricing. The second is the capacity market, an administrative overlay created to solve what economists call the "missing money" problem — the persistent gap between what generators earn selling energy and what they need to earn to justify the investment in new power plants. Together, these two market designs represent the central architecture of organized wholesale electricity markets in the United States, and the tension between them represents

one of the most consequential ongoing debates in energy policy.

* * *

6.1 Locational Marginal Pricing

6.1.1 The Price of a Megawatt-Hour at a Point in Space

At 5:00 p.m. on a sweltering August afternoon, the wholesale price of electricity in Midtown Manhattan might be $250 per megawatt-hour. At that exact same moment, 200 miles to the north in the rural Mohawk Valley, the price might be $35 per megawatt-hour. The electrons are indistinguishable. The product is identical. The moment is the same. Yet the prices differ by a factor of seven.

This is not a market failure. It is a market working precisely as designed. The mechanism responsible is called Locational Marginal Pricing, or LMP, and it is the intellectual foundation upon which all organized wholesale electricity markets in the United States are built. Developed in its modern form by the economist William Hogan at Harvard University and first implemented by PJM Interconnection in 1998, LMP solved a problem that had bedeviled electricity market designers from the beginning: how to incorporate the physical constraints of the transmission network into the price of energy itself.

The insight behind LMP is elegantly simple. Because electricity flows according to the laws of physics — not according to the contracts humans write — the cost of delivering power to any given location depends not only on the cost of generating it, but also on the path it must take through the transmission system and whether that system has the capacity to carry it. A cheap generator 500 miles away is useless if the transmission lines between that generator and the load are already at their thermal limits. In that situation, a more expensive local generator must be dispatched instead, and the true cost of serving that load is the cost of that more expensive generator — not the cheaper one that physics will not allow to deliver its power.

LMP captures this reality by assigning a unique price to every node — every bus, every injection point, every withdrawal point — in the transmission network. In PJM, there are more than 10,000 such nodes. In the Electric Reliability Council of Texas, there are roughly 4,000. Each node has its own price, recalculated every five minutes, reflecting the actual cost of serving one additional megawatt-hour of load at that precise location at that precise moment.

6.1.2 The Three Components of LMP

Every locational marginal price can be decomposed into three distinct components, each capturing a different aspect of the cost of delivering energy to a specific point on the grid.

The first component is the **system energy price**. This is the marginal cost of generating one additional megawatt-hour of electricity somewhere on the system, absent any transmission constraints. If the entire grid were a single copper plate — a network with infinite transmission capacity and no losses — every node would have the same price, and that price would equal the marginal cost of the most expensive generator needed to serve the system's total load. This hypothetical single-price world represents the energy component.

The second component is the **congestion cost**. When transmission lines reach their limits, the system operator can no longer dispatch generators in pure economic merit order. Instead, cheaper generators on the constrained side of the bottleneck must be curtailed, and more expensive generators on the load side must be dispatched in their place. The price difference this creates between the two sides of the constraint is the congestion component. It is positive at nodes downstream of the constraint (where load is competing for scarce transmission capacity) and negative at nodes upstream (where generation is abundant but cannot reach the load). Congestion costs are the market's way of signaling the economic value of transmission capacity — and, by extension, the economic value of building new transmission infrastructure or siting new generation in import-constrained areas.

The third component is **marginal losses**. As electricity flows through conductors, some energy is lost as heat — a consequence of the resistance of the wire itself. These losses are not uniform across the system; they increase with the square of the current flowing through a line and with the distance the power must travel. Marginal losses at any given node reflect the incremental increase in system-wide losses caused by serving one additional megawatt of load at that location. Nodes far from generators tend to have higher loss components; nodes close to generation tend to have lower or even negative loss components.

Together, these three components fully characterize the cost of energy at every point on the grid:

> LMP at Node *i* = System Energy Price + Congestion Component at *i* + Loss Component at *i*

When there is no transmission congestion and losses are neglected, every node has the same price. The moment a transmission line binds — the moment it reaches its thermal, voltage, or stability limit — prices separate, and the geography of the grid begins to matter.

Figure 6.2: The Merit Order Stack — How Electricity Prices Are Set (Source: EIA, Lazard LCOE, FERC Energy Primer)

6.1.3 Security-Constrained Economic Dispatch

The prices described above do not emerge from bilateral negotiations or open outcry trading floors. They are computed by an optimization algorithm known as Security-Constrained Economic Dispatch, or SCED, run by the Independent System Operator or Regional Transmission Organization responsible for the market.

SCED is, at its core, a linear programming problem. The objective function is straightforward: minimize the total cost of meeting system load. The decision variables are the output levels of every generator connected to the grid. The constraints are numerous and exacting. Each generator has a minimum and maximum output, a ramp rate that limits how quickly it can increase or decrease production, and a cost curve (derived from its fuel type, heat rate, and variable operating costs) that describes how the cost of production changes with output level. Each transmission line has a thermal rating that limits the power flow it can carry. And the entire solution must be feasible not only under current conditions but also under a set of contingency scenarios — the "security" in security-constrained — meaning that the system must remain operable even if any single major element (a large generator, a critical transmission line) suddenly fails.

The algorithm solves this problem simultaneously for the entire network, taking into account the physical power flow equations that govern how electricity distributes itself across parallel paths. The

solution yields two outputs: a dispatch instruction for every generator (how much power to produce) and a shadow price for every constraint. The shadow price on the system power balance constraint is the system energy price. The shadow price on each binding transmission constraint gives the congestion component. Together with the loss factors, these shadow prices constitute the LMPs at every node.

This computation is performed every five minutes in real-time markets and for every hour in day-ahead markets. It is one of the most complex optimization problems solved at commercial scale anywhere in the economy.

6.1.4 A Concrete Example: Price Formation in Practice

Consider a simplified two-node system to illustrate how LMP works in practice. Node A is in a rural area with abundant low-cost wind generation available at $20 per megawatt-hour. Node B is in a city with load of 1,000 megawatts and a local natural gas plant that can generate at $80 per megawatt-hour. A single transmission line connects the two nodes, with a capacity limit of 800 megawatts.

If the city's load is 700 megawatts, the cheap wind power from Node A can supply it entirely through the transmission line, which is well within its 800-megawatt limit. Both nodes price at approximately $20 per megawatt-hour — the system marginal cost. The congestion component is zero.

Now suppose the city's load rises to 1,000 megawatts. The transmission line can carry only 800 megawatts from Node A. The remaining 200 megawatts must come from the local gas plant at Node B. The LMP at Node B is now $80 per megawatt-hour — the cost of the marginal generator serving that location. The LMP at Node A remains at $20 per megawatt-hour. The $60 difference is the congestion component, reflecting the economic cost of the transmission bottleneck.

Notice what this accomplishes. The high price at Node B signals to the market that this is a valuable location for new generation investment. It signals to consumers in the city that electricity is expensive and demand reduction is valuable. It signals to transmission developers that relieving the constraint between Nodes A and B is worth up to $60 per megawatt-hour of incremental transfer capability. And it signals to the wind generators at Node A that, despite their low cost, they cannot capture the full value of their output because they are on the wrong side of a transmission constraint. Every one of these signals promotes economic efficiency.

6.1.5 Day-Ahead and Real-Time Markets

Organized wholesale electricity markets operate on two parallel timelines. The **day-ahead market** clears the day before delivery. Load-serving entities submit their forecast demand, generators submit their supply offers, and the ISO runs the SCED algorithm to produce a set of financially binding schedules and prices for each hour of the following day. The day-ahead market serves as the primary venue for forward procurement. It gives generators scheduling certainty, allows loads to lock in prices, and provides the system operator with a reliable plan for the next day's operations.

The **real-time market** operates during the actual delivery day, clearing every five minutes. It settles

the deviations between day-ahead schedules and actual real-time conditions. If load turns out to be higher than forecast, or a generator trips offline unexpectedly, the real-time market adjusts dispatch and produces prices that reflect the true marginal cost of meeting the system's needs in that moment. Real-time prices tend to be more volatile than day-ahead prices, spiking sharply during unexpected events and occasionally going negative when surplus renewable generation cannot be absorbed.

The two markets are linked by arbitrage. If traders expect real-time prices to exceed day-ahead prices, they will buy in the day-ahead market and sell in real time, pushing day-ahead prices up. If they expect the reverse, they will sell day-ahead and buy in real time, pushing day-ahead prices down. This arbitrage, conducted through instruments known as **virtual bids** (or "increment offers" and "decrement bids"), promotes convergence between the two markets. Virtual bidders perform a valuable economic function: they inject private information about future supply and demand conditions into day-ahead prices, improving the efficiency of unit commitment decisions and reducing the need for costly out-of-market interventions by the system operator.

6.1.6 Financial Transmission Rights

If LMP creates price differences between nodes, it also creates financial risk. A generator located at a low-priced node who has contracted to sell power at a high-priced node faces exposure to congestion costs. A load-serving entity purchasing power from a distant generator faces the same risk. The wider the congestion spread, the greater the financial exposure.

Financial Transmission Rights (FTRs), known in some markets as **Congestion Revenue Rights** (CRRs), are the primary hedging instrument for this risk. An FTR is a financial contract that entitles its holder to receive (or obligates the holder to pay) the congestion component of the price difference between two specified nodes, for a specified quantity of power over a specified time period. A holder of a 100-megawatt FTR from Node A to Node B would receive 100 times the hourly congestion price difference between the two nodes — providing a perfect hedge against congestion costs for a 100-megawatt transaction flowing from A to B.

FTRs are funded by the congestion revenues that the ISO collects through the settlement process. When transmission is congested, the ISO collects more from loads at high-priced nodes than it pays to generators at low-priced nodes. This surplus — the congestion rent — funds FTR payments. In a perfectly modeled system, the total congestion revenue exactly equals the total FTR obligation. In practice, modeling approximations, transmission outages, and changes in grid topology can cause congestion revenue to fall short, leading to FTR revenue inadequacy — a recurring challenge in several organized markets.

FTRs are allocated through annual and monthly auctions conducted by the ISO, and they can be traded in secondary markets. They serve multiple purposes: hedging commercial transactions, enabling efficient risk management, and providing long-term price signals about the value of transmission capacity on specific paths.

* * *

6.2 The "Missing Money" Problem

6.2.1 The Theoretical Foundation

The energy market described above is, in many respects, a triumph of economic design. It dispatches generators efficiently, manages transmission congestion through price signals rather than administrative rationing, and reveals the true marginal cost of electricity at every point on the grid. But it has a problem — one that strikes at the very heart of resource adequacy and long-term reliability.

The problem is this: even a perfectly competitive energy market, operating with complete efficiency in the short run, may not produce enough revenue to sustain the generation fleet needed for long-term reliability.

To understand why, consider the economics of a natural gas peaking plant — the kind of simple-cycle combustion turbine that runs only during the highest-demand hours of the year. Such a plant might cost $80,000 to $120,000 per megawatt to build. It has low fixed operating costs but high variable costs, typically in the range of $80 to $150 per megawatt-hour depending on gas prices and heat rate. Its revenue model depends not on running many hours at a modest margin, but on running very few hours at an enormous margin — selling power during those handful of extreme hours each year when demand is so high that prices spike far above normal levels.

In a truly uncapped market, those price spikes could, in theory, provide enough revenue to cover the fixed costs of the plant. During a heat wave or a polar vortex, when the system is at its absolute limit, the marginal cost of unserved load — the Value of Lost Load, or VOLL — could be extraordinarily high. Estimates of VOLL range from $5,000 to $35,000 per megawatt-hour, reflecting the enormous economic damage caused by involuntary blackouts. If energy prices were allowed to reach these levels, even a peaker that runs only 50 or 100 hours per year might earn enough scarcity rents to justify its construction.

But in practice, energy prices are capped. Every organized market in the United States imposes some form of administrative price ceiling on energy offers. These caps exist for understandable reasons. Wholesale electricity markets are highly concentrated, especially during scarcity conditions, when only a few generators remain available to serve load. The potential for the exercise of market power — for generators to withhold output or inflate their offers to extract monopoly rents — is greatest precisely when prices need to be highest. Regulators, scarred by the California energy crisis of 2000-2001, have been reluctant to allow unconstrained pricing during scarcity events.

The consequence is what economists have termed the **"missing money" problem**. Administrative price caps, offer mitigation rules, and the general reluctance of regulators to permit sustained price spikes

suppress the scarcity rents that generators would need to earn in an uncapped market. The revenue that generators receive from selling energy alone falls short — sometimes dramatically — of the revenue needed to cover their all-in costs, including capital recovery and a reasonable return on investment.

6.2.2 Quantifying the Gap

The magnitude of the missing money varies by market, by year, and by generator type, but the pattern is persistent. Studies by Potomac Economics, the market monitor for several ISOs, have repeatedly documented that energy and ancillary service revenues in markets like NYISO, MISO, and SPP fall well short of the annualized cost of new entry for combustion turbines — the benchmark technology typically used to assess resource adequacy economics.

The problem is particularly acute for resources that are needed for reliability but operate infrequently. A peaking plant that runs only during the 20 to 100 highest-load hours per year earns almost all of its energy market revenue during those hours. If prices during those hours are capped at $1,000 or $2,000 per megawatt-hour rather than being allowed to reach the VOLL, the plant's annual revenue can fall tens of thousands of dollars per megawatt short of what is needed to justify new investment.

The specific cap levels vary by market. PJM and most eastern markets have historically imposed offer caps in the range of $1,000 to $2,000 per megawatt-hour, though these have been increased over time. ERCOT, the Texas market, has adopted the most aggressive approach to scarcity pricing, raising its system-wide offer cap to $9,000 per megawatt-hour — a level intended to represent a reasonable approximation of VOLL. ERCOT's designers explicitly chose this path as an alternative to a capacity market, betting that sufficiently high scarcity prices could solve the missing money problem without administrative procurement of capacity.

6.2.3 The Reliability Externality

The missing money problem is more than an abstract concern about investor returns. It has a direct and dangerous implication for reliability. If generators cannot earn enough revenue to cover their costs, they will eventually retire. If new generators cannot expect to earn enough revenue to justify their construction, they will not be built. Over time, the reserve margin — the cushion of available generation above peak demand — will erode. Eventually, it will erode far enough that involuntary load shedding becomes inevitable.

The difficulty is that reliability is a public good, or more precisely, it is a good with a significant positive externality. When a generator is available to produce power during a system emergency, it benefits not only the customers who consume its output but all customers on the system, because its availability reduces the probability of a cascading blackout that would affect everyone. Individual generators, making private investment decisions based on expected market revenues, do not internalize this reliability externality. The result is a systematic underinvestment in generation capacity relative to

the socially optimal level.

This is the theoretical justification for capacity markets. If energy markets, even well-designed ones, cannot produce revenues sufficient to maintain the generation fleet needed for reliable service — and if the consequence of underinvestment is not merely higher prices but the physical failure of the power system — then some additional mechanism is needed to close the gap.

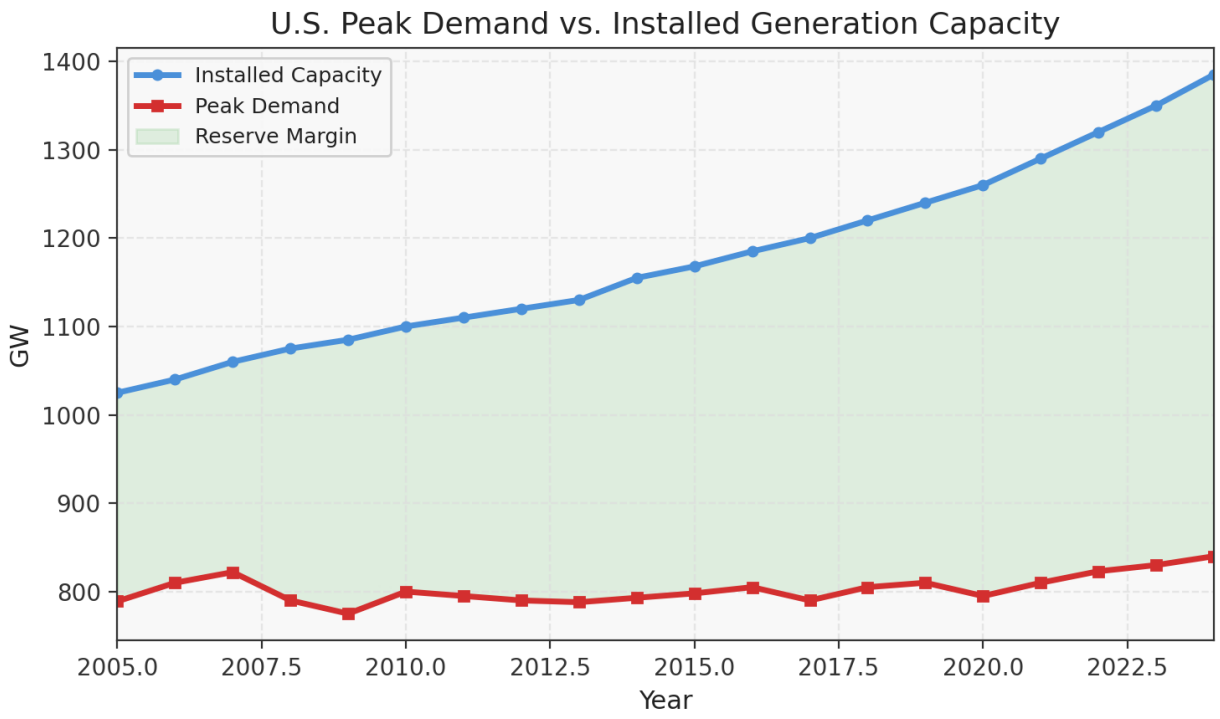

Figure 6.1: U.S. Peak Demand vs. Installed Generation Capacity (Source: EIA Electric Power Annual, NERC Long-Term Reliability Assessment)

6.2.4 The Energy-Only Alternative: Scarcity Pricing and Demand Response

Not everyone agrees that capacity markets are the right solution. A vocal school of thought, well-represented among market economists and particularly influential in Texas, argues that the missing money problem is itself a product of regulatory intervention. If regulators would simply remove price caps, eliminate offer mitigation during scarcity, and allow prices to rise to VOLL during genuine shortages, the resulting scarcity rents would provide sufficient revenue to sustain the generation fleet. No capacity market would be needed.

ERCOT has been the primary laboratory for this "energy-only" philosophy. With its $9,000 per megawatt-hour price cap and its Operating Reserve Demand Curve (ORDC) — an administrative mechanism that adds a scarcity premium to energy prices when operating reserves fall below target levels — ERCOT has attempted to create the conditions under which energy prices alone can support resource adequacy. The ORDC is particularly innovative: rather than waiting for reserves to drop to zero

and triggering emergency procedures, it begins adding a graduated premium to energy prices as reserves decline, creating a continuous price signal that rises as the system approaches scarcity.

Proponents of the energy-only approach also emphasize the role of **demand response** — the ability of consumers to reduce their electricity consumption in response to high prices. In a world with effective demand response, prices need not rise to astronomically high levels during scarcity events, because demand reduction effectively increases the available supply margin. Smart thermostats, interruptible industrial loads, and behind-the-meter battery storage can all function as demand response resources, flattening peak demand and reducing the need for rarely used peaking capacity.

The counter-argument is that the energy-only approach requires a tolerance for price volatility and supply risk that may be politically unsustainable. When ERCOT's system collapsed during Winter Storm Uri in February 2021 — with prices pinned at $9,000 per megawatt-hour for days, retail customers facing catastrophic bills, and millions of Texans enduring extended blackouts — the political backlash was swift and severe. The Texas legislature subsequently imposed new reliability mandates and, in a notable irony, began exploring the creation of a capacity-like mechanism, suggesting that even the most committed energy-only market may eventually confront the limits of price signals alone.

* * *

6.3 PJM's Reliability Pricing Model

6.3.1 Origins and Rationale

PJM Interconnection, the regional transmission organization serving all or parts of 13 states and the District of Columbia, operates the largest competitive wholesale electricity market in the world. It is also home to the most developed and controversial capacity market in the United States: the Reliability Pricing Model, or RPM.

RPM was established in 2007 to replace PJM's earlier capacity construct, which had devolved into a system of bilateral contracts at administratively determined prices. The earlier system had failed to attract new generation investment, and PJM's reserve margins were declining toward uncomfortable levels. RPM was designed to create a forward-looking, market-based mechanism for procuring the generation capacity needed to maintain reliability — a mechanism that would produce transparent prices, send locational investment signals, and provide revenue certainty to investors contemplating new power plant construction.

Since its inception, RPM has cleared hundreds of billions of dollars in capacity obligations, shaped investment decisions across the PJM footprint, and generated sustained controversy about its design, its costs, and its effectiveness. Understanding RPM in detail is essential for understanding how American

electricity markets actually work.

6.3.2 The Base Residual Auction

The centerpiece of RPM is the **Base Residual Auction** (BRA), a centralized auction held approximately three years before the delivery year. (The lead time has varied over RPM's history; PJM has shifted between three-year and shorter forward periods, but the three-year-ahead structure represents the core design intent.) The three-year forward period serves a specific purpose: it provides developers with enough time to permit, finance, and construct a new power plant in response to a capacity price signal.

In the BRA, PJM determines how much capacity the system will need during the delivery year — based on peak load forecasts, target reserve margins, and the installed reserve margin determined through probabilistic reliability studies — and then solicits offers from all resources capable of providing that capacity. Resources include existing generators, planned new generators, demand response providers, energy efficiency programs, and in recent years, storage resources. Each resource submits an offer price reflecting the minimum payment it would accept to take on a capacity obligation for the delivery year.

On the demand side, PJM does not submit a single fixed quantity. Instead, it constructs a downward-sloping **demand curve** — the Variable Resource Requirement (VRR) curve — that expresses the system's willingness to pay for capacity as a function of the reserve margin. The VRR curve is anchored to a concept known as the **Net Cost of New Entry** (Net CONE), which represents the annualized cost of building a new peaking generator (the reference technology, typically a combustion turbine or combined-cycle plant) minus the expected energy and ancillary service revenues that plant would earn in its first year of operation. Net CONE thus represents the "missing money" itself — the revenue gap that the capacity market must fill.

The VRR curve is constructed so that the clearing price equals Net CONE when the reserve margin exactly equals the target reserve margin. If the reserve margin falls below target (indicating tighter supply), the curve produces a price above Net CONE, providing a stronger investment signal. If the reserve margin exceeds the target (indicating surplus capacity), the curve produces a price below Net CONE, discouraging further entry. This sloped demand curve is one of RPM's most important design features: it provides price stability relative to a vertical demand curve (where small changes in supply or demand can produce enormous price swings) while still providing meaningful price signals about the system's capacity position.

6.3.3 Supply Offers and the Clearing Process

On the supply side, the auction operates as a descending-clock mechanism, though in practice the clearing is determined computationally. Each resource submits an offer specifying the price at which it is willing to accept a capacity obligation. For existing generators, the offer typically reflects the going-forward costs that the resource must cover to remain operational — fixed operation and maintenance costs, property taxes, insurance, and any capital expenditures needed to maintain reliability.

For new generators, the offer reflects the full annualized cost of construction, net of expected energy revenues.

PJM stacks the supply offers from lowest to highest and intersects them with the VRR demand curve. The clearing price is determined by the intersection point: all resources that offered at or below the clearing price receive a capacity commitment and are paid the clearing price. Resources that offered above the clearing price are not selected and receive no capacity payment. This is a uniform-price auction — all cleared resources receive the same price, regardless of their individual offer level — which mimics the marginal pricing logic of the energy market.

The auction clears not only for the PJM system as a whole but also for defined sub-regions known as **Locational Deliverability Areas** (LDAs). LDAs correspond to transmission-constrained load pockets where local capacity has particular value because imports may be limited during peak conditions. If a particular LDA faces a binding transmission constraint in the capacity auction, it can clear at a higher price than the rest of the system, signaling the need for new local generation or demand-side resources. This locational price separation mirrors, in the capacity context, the congestion-driven price separation that LMP produces in the energy market.

6.3.4 Capacity Performance and Non-Performance Penalties

In its original form, RPM required capacity resources only to be available during summer peak hours, and penalties for non-performance were modest. This changed dramatically after the polar vortex events of January 2014, when extreme cold weather caused widespread generator failures across PJM. Approximately 22 percent of PJM's generation capacity experienced forced outages during the event, primarily due to frozen fuel supply lines, equipment malfunctions in extreme cold, and inadequate winterization. The near-miss with rolling blackouts across a region serving 65 million people exposed a fundamental weakness in the capacity market's design: resources were being paid to be available, but many were not actually performing when they were needed most.

PJM responded by introducing **Capacity Performance** (CP) requirements, effective for the 2020/2021 delivery year. Under CP, all capacity resources must be available year-round (not just during summer peaks) and face severe financial penalties for failure to perform during declared emergency events. The non-performance charge rate is set at the Net CONE value divided by the expected number of Performance Assessment Hours — the hours during which the system is under stress and performance is measured. This penalty structure can impose charges of several thousand dollars per megawatt-hour of shortfall, creating a powerful financial incentive for generators to invest in fuel assurance, weatherization, and maintenance.

Correspondingly, resources that over-perform during emergencies — producing more than their capacity obligation — receive bonus payments funded by the non-performance penalties assessed on underperforming resources. This creates a zero-sum transfer within the capacity market: reliable resources are rewarded at the expense of unreliable ones, strengthening the link between capacity payments and actual performance.

The CP construct has been controversial. It has increased costs for intermittent resources like wind and solar, which cannot guarantee availability during all hours. It has imposed significant financial risk on generators, particularly older coal and gas plants with less predictable forced outage rates. And it has raised questions about the appropriate balance between penalty severity and the commercial risk that generators are willing to bear.

6.3.5 The Net Cost of New Entry

The Net CONE is perhaps the single most consequential parameter in RPM's design, because it anchors the demand curve and thus determines the price level around which the auction clears. Getting Net CONE right is essential; getting it wrong can either over-procure capacity at excessive cost to consumers or under-procure capacity and jeopardize reliability.

Calculating Net CONE requires estimating the all-in annualized capital cost of a new reference technology (including construction costs, financing costs, property taxes, insurance, and a return on equity), and then subtracting the estimated energy and ancillary service revenues that such a plant would earn in the PJM market. Both components are subject to considerable uncertainty. Construction costs fluctuate with commodity prices, labor markets, and interest rates. Energy revenue forecasts depend on assumptions about future fuel prices, load growth, renewable penetration, and environmental regulations.

PJM periodically reviews and updates its Net CONE estimates through a stakeholder process, and the results are often contentious. Generators typically argue that Net CONE is set too low, suppressing capacity prices and discouraging investment. Consumer advocates argue that Net CONE is set too high, inflating capacity costs and enriching incumbent generators. The reference technology itself is debated: should it be a simple-cycle combustion turbine, a combined-cycle plant, or some other technology? Each choice produces a different Net CONE and, consequently, different auction outcomes.

6.3.6 The Minimum Offer Price Rule Controversy

No aspect of RPM has generated more controversy than the **Minimum Offer Price Rule** (MOPR). The MOPR was originally designed to prevent buyer-side market power — the ability of large load-serving entities to suppress capacity prices by subsidizing new generation that offers into the auction at below-market prices. If a state-subsidized generator offers into the capacity auction at zero or near-zero prices, it displaces unsubsidized generators that need the capacity revenue to survive. The result is lower auction clearing prices, which may drive existing generators into premature retirement and ultimately undermine reliability.

The MOPR addresses this by imposing a price floor on capacity offers from new resources, typically set at or near the estimated cost of new entry for that resource type. If a new generator's offer falls below the MOPR floor, its offer is reset to the floor price, preventing it from artificially suppressing the auction clearing price.

The controversy erupted as states increasingly began using subsidies to advance clean energy

policies. When states like New Jersey, Illinois, and Ohio created Zero Emission Credit (ZEC) programs to support struggling nuclear plants, or when states provided renewable portfolio standard (RPS) mandates and associated renewable energy credits that effectively subsidized wind and solar development, the MOPR became a flashpoint in a constitutional-scale conflict between state energy policy and federal market regulation.

At its most expansive, FERC Order in 2019 extended the MOPR to apply to virtually all state-subsidized resources, potentially blocking subsidized renewables, nuclear plants, and demand response programs from clearing the PJM capacity auction. Critics argued that this effectively gave FERC veto power over state energy policy — that by preventing subsidized clean energy resources from competing in the capacity market, the MOPR forced consumers to pay twice: once for the state-mandated clean energy resource and again for the fossil-fuel resource that cleared the capacity market in its place.

In 2021, FERC reversed course and effectively eliminated the broad MOPR, replacing it with a more limited set of market power screens. The policy pendulum on MOPR has swung repeatedly and may swing again, reflecting the fundamental unresolved tension between competitive wholesale markets administered under federal jurisdiction and state energy policies that increasingly seek to direct the generation mix toward specific public policy outcomes.

6.3.7 Comparative Design: ISO-NE and NYISO

PJM's RPM is not the only capacity market in the United States, and comparing it with the designs adopted by other ISOs illuminates the range of approaches to solving the missing money problem.

ISO New England's Forward Capacity Market (FCM) shares many structural features with RPM — a forward auction, a sloped demand curve, and locational pricing for constrained zones — but differs in several important respects. The FCM's forward period is approximately three and a half years, slightly longer than PJM's. ISO-NE has also experimented with a "pay-for-performance" penalty structure that preceded and in some ways inspired PJM's Capacity Performance reforms. The FCM has grappled with its own version of the MOPR debate, as New England states have aggressively pursued offshore wind and other clean energy resources that interact uneasily with the capacity market's pricing mechanisms. ISO-NE has adopted a "Substitution Auction" approach that allows subsidized resources to replace existing resources' capacity obligations, partially mitigating the two-payment problem that plagued the expansive MOPR.

NYISO's capacity market differs more substantially. Rather than a single forward auction, NYISO operates a series of monthly and seasonal capacity auctions, with capacity obligations traded on a shorter-term basis. NYISO divides its system into capacity zones — including the highly constrained New York City and Long Island zones, where local capacity commands a significant premium — and clears separate demand curves for each zone. The NYISO demand curve is anchored to Net CONE estimates specific to each zone, reflecting the dramatically different costs of building new generation in Midtown Manhattan versus rural upstate New York. NYISO's approach provides less long-term price certainty than PJM's three-year-ahead auction, but it also offers more flexibility to adapt to rapidly

changing market conditions and state policy directives.

The variation among these three designs — all operating within the same FERC-jurisdictional framework, all attempting to solve the same missing money problem — underscores that there is no single "correct" capacity market architecture. Each design reflects a particular set of judgments about the appropriate balance between forward certainty and near-term flexibility, between market competition and administrative determination, and between federal market design and state policy autonomy.

* * *

6.4 The Enduring Tension

The relationship between energy markets and capacity markets is often characterized as complementary — two halves of a complete revenue stream for generators, one compensating for what the other cannot provide. But the relationship is also one of deep tension. Every dollar that a generator earns in the capacity market is, in a sense, an admission that the energy market has failed to provide adequate compensation on its own. And every increase in capacity payments raises the question of whether the energy market's price signals are being systematically undermined.

This tension manifests in practical ways. When capacity markets clear at high prices, they can sustain generators that would otherwise retire in the face of low energy prices — keeping the generation fleet larger (and the reserve margin higher) than energy market signals alone would support. But this surplus capacity, in turn, suppresses energy prices further, widening the missing money gap and increasing the system's dependence on capacity payments. The dynamic is self-reinforcing: the more capacity markets pay, the less energy markets pay, and the more capacity markets need to pay.

The entrance of renewable energy at scale has intensified this dynamic. Wind and solar resources have near-zero marginal costs; when they are producing, they suppress energy prices for everyone, reducing the energy market revenues available to conventional generators. But wind and solar are intermittent — their availability is determined by weather, not by market need. The capacity value of a wind farm is substantially less than its nameplate capacity, because it cannot guarantee production during peak demand hours. The result is a system that needs conventional generators for reliability but cannot pay them enough through energy prices alone to remain viable — the missing money problem, amplified by the clean energy transition.

No consensus has emerged on how to resolve this tension. Some economists advocate for more aggressive scarcity pricing in energy markets, reducing the need for capacity payments. Others argue for fundamental reforms to capacity market design, including multi-year capacity commitments, technology-specific procurement targets, or integration of clean energy attributes into the capacity auction itself. Still others contend that the entire construct of competitive wholesale markets is poorly suited to a power system dominated by capital-intensive, zero-marginal-cost renewable resources, and

that some form of long-term central procurement — closer to the old utility planning model — may ultimately prove necessary.

What is clear is that the design of energy and capacity markets is not a technical exercise that can be completed once and left alone. It is an ongoing process of institutional adaptation, responding to changes in technology, policy, fuel markets, and the physical climate. The markets described in this chapter are not static architectures; they are living systems, continuously revised and reformed in response to the evolving challenge of keeping the most complex machine ever built by human civilization running reliably, affordably, and — increasingly — cleanly.

* * *

Chapter Summary

This chapter has examined the two principal market mechanisms that govern wholesale electricity procurement in the organized markets of the United States. Locational Marginal Pricing provides the short-run price signal, reflecting the true cost of delivering energy to each point on the transmission network, incorporating the effects of generation costs, transmission congestion, and electrical losses. The missing money problem reveals the limitation of short-run marginal cost pricing for a commodity that requires enormous fixed capital investment: prices that are efficient in the short run may be insufficient in the long run to sustain the generation fleet needed for reliability. Capacity markets, exemplified by PJM's Reliability Pricing Model, represent the dominant institutional response to this problem — an administrative mechanism layered atop the energy market to procure and compensate the resources needed to maintain target reserve margins.

The chapter has shown that these market structures, while analytically separable, are deeply interdependent. Energy prices influence capacity prices and vice versa. State policies interact with federal market design in ways that are economically consequential and legally contested. And the accelerating transition to a low-carbon grid is straining both market structures in ways their designers did not fully anticipate. The next chapter turns to the governance institutions — FERC, the state commissions, the ISOs themselves — that oversee these markets and manage the conflicts that arise from their inherent tensions.

* * *

Part IV

Regional Profiles and Comparative Analysis

Chapter 7: ERCOT: The Texas Experiment

Introduction

No state in the American union has pursued a more distinctive path in electricity policy than Texas. While the rest of the continental United States operates within two vast synchronized alternating-current networks — the Eastern Interconnection and the Western Interconnection — Texas has, with characteristic independence, chosen to go it alone. The Electric Reliability Council of Texas, known universally as ERCOT, manages a grid that is, by deliberate design, an electrical island. It is connected to its neighbors only through a handful of small direct-current ties that are carefully engineered to avoid the legal threshold that would bring the system under federal regulatory authority.

This is not merely a geographic or engineering curiosity. It is a foundational policy choice that has shaped every dimension of how electricity is generated, priced, traded, and governed across a territory that, if Texas were an independent nation, would rank among the ten largest electricity-consuming economies in the world. ERCOT serves roughly 27 million customers across approximately 90 percent of the state's electric load, managing a system with a peak demand that has exceeded 85 gigawatts in summer — a figure larger than the peak demand of most European countries.

The Texas experiment rests on two interconnected premises. The first is jurisdictional: by remaining electrically isolated from interstate commerce, Texas preserves the authority of its own regulatory institutions — principally the Public Utility Commission of Texas — to design and govern its electricity markets without deference to the Federal Energy Regulatory Commission. The second premise is economic: within that zone of regulatory autonomy, Texas has constructed an energy-only market, a design that rejects the capacity payments and resource adequacy mandates common in other organized wholesale markets. Instead, ERCOT relies on scarcity pricing — the theory that allowing wholesale electricity prices to spike to extraordinary levels during periods of genuine shortage will, over time, send sufficient investment signals to attract the generation capacity the system requires.

For roughly two decades, this approach was widely celebrated. Texas attracted enormous investment in both natural gas and wind generation. Wholesale prices were, on average, among the lowest in the

nation. The state's electricity sector was held up as a model of deregulation done right — proof that competitive markets, liberated from heavy-handed federal oversight, could deliver reliable and affordable power.

Then came February 2021.

Winter Storm Uri did not merely expose operational failures in the Texas grid. It posed a fundamental challenge to the intellectual architecture of the Texas experiment itself, raising questions that the state — and the nation — continue to grapple with years later. This chapter examines both the design and the stress test: the economics of ERCOT's energy-only market, the strategic logic and trade-offs of jurisdictional isolation, and what happens when the assumptions underlying both are overwhelmed by physical reality.

* * *

I. Energy-Only Economics: Scarcity Pricing and the "Laissez-Faire" Grid

The Philosophical Foundation

To understand ERCOT's market design, one must first understand what it rejected. In most organized wholesale electricity markets across the United States — PJM, ISO New England, NYISO, and MISO among them — system operators have concluded that energy market revenues alone are insufficient to ensure that enough generation capacity will be built and maintained to meet future demand with an acceptable margin of reliability. These markets therefore layer a capacity market or capacity obligation on top of the energy market. Generators receive payments not only for the electricity they produce but also for their commitment to be available to produce electricity when called upon. Load-serving entities are required to procure capacity sufficient to cover their projected peak demand plus a reserve margin, and the cost of those capacity commitments is passed through to consumers.

Texas looked at this model and said no.

The architects of the ERCOT market, which launched in its current nodal form in 2010 after operating in a zonal configuration since 2001, made a deliberate philosophical and economic choice. They argued that capacity markets distort investment signals, suppress energy prices below efficient levels, and create a regulatory apparatus that inevitably drifts toward central planning. In their view, the only honest price signal is the price of energy itself. If the market is allowed to function — if prices are permitted to rise high enough during periods of scarcity to reflect the true marginal value of electricity when supply is short — then entrepreneurs and investors will respond rationally. They will build the

generators, demand-response resources, and storage facilities that the system needs, attracted by the prospect of earning very high revenues during those critical hours.

This is the energy-only market in its purest form. It is, in essence, a bet on the price mechanism. It asks market participants to accept a revenue pattern that resembles, in the memorable phrase used by many analysts, "feast or famine" — long stretches of moderate or even below-cost prices punctuated by brief episodes of extraordinary earnings.

The Mechanics of the Offer Cap

Central to the functioning of any energy-only market is the level of the offer cap — the maximum price at which generators can offer their output into the wholesale market. The offer cap serves a dual purpose. It defines the ceiling for scarcity revenues, and it communicates to investors the maximum earning potential during shortage conditions.

When ERCOT's competitive market was first established, the system-wide offer cap was set at $1,000 per megawatt-hour. This figure was quickly recognized as too low to sustain adequate investment. If prices could only reach $1,000/MWh during the most extreme conditions, the total scarcity revenue earned by peaking units over the course of a year would fall short of what was needed to justify the capital cost of building and maintaining those units. The cap was subsequently raised in a series of steps — to $3,000/MWh, then to $4,500/MWh, then to $7,000/MWh, and ultimately to $9,000/MWh. Each increase reflected a judgment by the PUCT that the existing cap was suppressing the investment signal.

The $9,000/MWh figure is worth pausing over. It is roughly 100 to 200 times the average wholesale price of electricity in ERCOT during normal conditions. A generator earning $9,000/MWh for even a few hours can earn more revenue in that brief window than it would in weeks of operation at normal prices. The logic is that the prospect of such returns — however infrequent — should be sufficient to incentivize the construction of peaking generation capacity, which by definition operates only during the hours of highest demand and tightest supply.

This is an elegant theory. Its practical challenge is that it requires market participants — and, critically, their lenders and investors — to tolerate enormous revenue volatility and to make capital allocation decisions based on probabilistic expectations about events that may occur only a handful of hours per year.

The Operating Reserve Demand Curve

Recognizing that the offer cap alone was an imperfect tool, the PUCT and ERCOT introduced a more sophisticated scarcity pricing mechanism in 2014: the Operating Reserve Demand Curve, or ORDC. The ORDC represents one of the more intellectually ambitious elements of ERCOT's market design, and understanding it is essential to understanding how the energy-only model is meant to function in practice.

The ORDC works as follows. In real time, ERCOT continuously calculates the probability that

available operating reserves — the cushion of unused generation capacity standing ready to respond to sudden changes in supply or demand — will fall below the minimum level required to maintain system reliability. This minimum level is defined as 2,000 megawatts, a threshold below which the risk of involuntary load shedding (rolling blackouts) becomes unacceptably high.

As actual reserves decline toward and below this threshold, the ORDC assigns an increasing price adder that is added to the real-time settlement price for energy. The adder is calculated using a statistical model that estimates the Loss of Load Probability — the likelihood that reserves will be insufficient to prevent load shedding — at each reserve level. The adder is, conceptually, the product of the probability of shortage and the Value of Lost Load (VOLL), which is set administratively. In ERCOT's implementation, the VOLL was set equal to the system-wide offer cap of $9,000/MWh.

The mathematical elegance of this mechanism lies in its attempt to price scarcity on a continuous curve rather than as a binary event. When reserves are abundant — say, 6,000 megawatts or more — the ORDC adder is negligible. As reserves tighten to 4,000 MW, the adder begins to become noticeable. As reserves fall below 3,000 MW, the adder rises steeply. And if reserves approach the 2,000 MW minimum contingency level, the adder can reach thousands of dollars per megawatt-hour.

The ORDC thus functions as an administrative scarcity pricing mechanism embedded within the real-time energy market. It increases energy prices before involuntary curtailments actually occur, sending a price signal that is meant to incentivize several responses simultaneously: it encourages generators to remain online and maximize output; it signals demand-response resources to reduce consumption; and, over the longer term, it contributes to the revenue stream that is supposed to justify investment in new generation capacity.

One of the ORDC's key design virtues, from the perspective of its advocates, is that it prices reserves as a public good. In a traditional energy market without an ORDC, reserves provide reliability benefits to the entire system but are not directly compensated at their marginal value. The ORDC corrects this by creating a revenue stream that flows to all generators providing reserves, thereby internalizing what would otherwise be an externality.

The Peaker Net Margin and Investment Adequacy

How does one evaluate whether ERCOT's energy-only market is actually producing adequate investment signals? The PUCT and ERCOT's Independent Market Monitor have relied on a metric known as the Peaker Net Margin (PNM) to answer this question.

The Peaker Net Margin is conceptually straightforward. It measures the net revenue that a hypothetical new natural gas combustion turbine — the marginal peaking unit — would have earned in the ERCOT market over a given year, accounting for both energy and ancillary service revenues and subtracting variable operating costs. This net revenue figure is then compared to the annualized fixed cost of building and maintaining such a unit, often referred to as the Cost of New Entry, or CONE.

If the Peaker Net Margin consistently equals or exceeds the CONE, the market is, in theory, sending an adequate investment signal. A rational investor would build new peaking capacity because the

expected return justifies the capital expenditure. If the PNM falls persistently below the CONE, the market is failing to incentivize the entry of new generation needed to maintain reliability.

In practice, the PNM in ERCOT has been highly variable from year to year, reflecting the inherent volatility of the energy-only model. In years with extreme weather events — hot summers, cold winters, or both — the PNM has spiked to levels well above the CONE, driven by hours of very high scarcity pricing. In mild years, the PNM has fallen well short. Proponents of the energy-only model argue that it is the multi-year average of the PNM that matters, not any single year's result, and that the long-run average has been broadly adequate. Critics counter that investment decisions are not made on the basis of backward-looking averages; they are made by human beings and institutions that must commit capital under uncertainty and that may systematically underweight the probability of extreme price events, particularly if those events have not occurred recently.

This debate, it should be noted, is not merely academic. The feast-or-famine revenue pattern creates a distinctive set of incentive problems. A developer contemplating the construction of a new peaking plant must persuade lenders and equity investors that the plant will earn enough revenue in a small number of high-price hours to compensate for hundreds of hours of zero or near-zero margins. The lender must believe that the regulatory and political environment will allow prices to remain at $9,000/MWh when they occur — a belief that is tested every time high prices generate political controversy and consumer anger. If the market perceives even a modest probability that the offer cap will be lowered, or that some form of price mitigation will be imposed during the next crisis, the expected value of scarcity revenues declines, and the investment case weakens.

This is the "missing money" problem in an energy-only context. Unlike capacity markets, where the missing money problem is addressed through explicit capacity payments, ERCOT's design assumes that the energy market itself, properly structured, provides all the revenue needed. The ORDC is the principal mechanism for bridging any gap. Whether it succeeds is a question that has been debated continuously since its inception and that took on new urgency after the events of February 2021.

The Role of Renewables in Energy-Only Economics

An additional complication in ERCOT's energy-only market has been the rapid growth of wind and, more recently, solar generation. Texas has more installed wind capacity than any other state, and its solar capacity has grown exponentially since the late 2010s. These resources have near-zero marginal costs: once built, the fuel is free. Their effect on energy prices has been to push average wholesale prices lower, particularly during the windy overnight hours and sunny midday periods when renewable output is highest.

This dynamic, often called the "merit order effect," is not unique to ERCOT, but it is pronounced there because of the sheer scale of renewable penetration and the absence of a capacity market to provide a revenue backstop for conventional generators. As average energy prices decline under the influence of zero-marginal-cost renewables, the investment case for new dispatchable generation — the gas plants and other firm resources that are needed to fill in when the wind stops blowing and the sun stops shining

— becomes more dependent on ever-rarer episodes of high scarcity pricing. The feast becomes more concentrated; the famine, more prolonged.

This is one of the central tensions in ERCOT's energy-only design as it has evolved. The market was conceived in an era when the marginal generator was a natural gas plant. It now operates in a world where the marginal resource during many hours is a wind farm or solar array, and where the scarcity pricing mechanism must work harder to compensate dispatchable resources that run fewer hours but are no less essential to reliability.

* * *

II. The "Island" Strategy: How Texas Avoids Federal Jurisdiction

Historical and Legal Origins

The story of how Texas came to operate an electrically isolated grid is, like many aspects of Texas history, a story about sovereignty — or at least the fierce desire for it.

The Federal Power Act of 1935 granted the Federal Power Commission (the predecessor to the Federal Energy Regulatory Commission, or FERC) jurisdiction over the transmission and wholesale sale of electricity in interstate commerce. The key operative phrase is "in interstate commerce." If electricity does not cross state lines, the federal jurisdictional hook does not attach.

Texas utilities, alert to this principle from the very beginning, were careful to develop their transmission networks in a way that avoided synchronous interconnection with utilities in neighboring states. While the Eastern and Western Interconnections grew through decades of utilities linking their systems together for mutual reliability and economic benefit, Texas utilities maintained their separation. The state's transmission network was coordinated, beginning in the 1940s, through a succession of power pools and reliability organizations that eventually evolved into ERCOT, which was established in 1970 in the wake of the Northeast Blackout of 1965 and the subsequent push for greater coordination of reliability standards.

But coordination within Texas is not the same as interconnection with the rest of the country. ERCOT's boundaries were drawn to encompass the synchronous grid that remained electrically isolated from the Eastern and Western Interconnections. Utilities in the Texas Panhandle, the eastern fringe of the state, and the far western tip — areas that were historically interconnected with out-of-state systems — were excluded from ERCOT and remain under FERC jurisdiction through their membership in the Southwest Power Pool or the Western Interconnection.

The result is a legal architecture of considerable importance. Because ERCOT does not engage in interstate commerce in electricity, the PUCT exercises plenary regulatory authority over the wholesale

market, transmission planning, and retail competition within ERCOT's footprint. FERC has no jurisdiction over ERCOT's market rules, transmission rates, or market participant behavior. This means that the PUCT — a three-member body appointed by the Governor of Texas — has far greater latitude to design and modify market rules than regulators in any other state. It can raise or lower the offer cap, modify the ORDC, impose new market mechanisms, or restructure the market entirely without seeking FERC approval or navigating the complex stakeholder processes that characterize FERC-jurisdictional markets.

The Governance Structure of ERCOT

ERCOT itself occupies an unusual position in the American electricity landscape. It is an independent system operator — an ISO — but it is not a FERC-jurisdictional ISO like PJM, MISO, or the California ISO. It operates under the oversight of the PUCT and the Texas Legislature, not under a FERC-approved tariff.

ERCOT is organized as a membership-based 501(c)(4) nonprofit corporation. Its board of directors has, over the years, undergone significant restructuring, particularly in the aftermath of Winter Storm Uri. Prior to 2021, the board included a mix of market participant representatives, consumer representatives, and independent members. The post-Uri legislative reforms significantly altered this structure, shifting toward a board composed primarily of members selected by the state's political leadership, reflecting a judgment that the pre-existing governance model had been insufficiently accountable to the public interest.

ERCOT's operational responsibilities mirror those of other ISOs: it manages the real-time dispatch of generation, operates the day-ahead and real-time energy markets, coordinates transmission planning, and administers the ancillary services markets that procure the operating reserves necessary for reliability. But it does so within a governance framework that is, by design, more directly responsive to state-level political authority than its counterparts elsewhere in the country.

The Trade-Offs of Isolation

The benefits of ERCOT's island strategy are real but are accompanied by significant costs that become most apparent during emergencies.

On the benefit side, regulatory independence has allowed Texas to move more quickly and with greater flexibility in market design than FERC-jurisdictional regions. The Texas Legislature's decision to restructure the electricity market and introduce retail competition, enacted through Senate Bill 7 in 1999, was implemented entirely under state authority. The PUCT's ability to modify the ORDC, adjust the offer cap, or introduce new market products does not require FERC proceedings that can take years to resolve. Market design changes that would require a complex Section 205 or Section 206 filing under the Federal Power Act can be implemented in ERCOT through PUCT rulemaking on a much shorter timeline.

The cost, however, is physical isolation. ERCOT has only two direct-current (DC) ties to the Eastern Interconnection — small-capacity links that can transfer a few hundred megawatts — and two DC ties to Mexico. These ties are deliberately limited in capacity to avoid the argument that ERCOT is engaged in interstate (or international) commerce on a scale that would trigger FERC jurisdiction. The legal theory underlying ERCOT's jurisdictional independence has never been definitively tested in court, and there is reason to believe that the federal government could assert jurisdiction if it chose to press the issue. But a tacit understanding has prevailed for decades: Texas does not push electricity across state lines, and the federal government does not push jurisdiction into Texas.

The practical consequence is that when ERCOT faces a supply emergency — whether from extreme weather, widespread generation outages, or unexpected demand — it cannot draw on the vast resources of the Eastern or Western Interconnection in the way that, say, a utility in Ohio or Georgia can. An emergency in PJM can be partially alleviated by imports from MISO, NYISO, or the Southeast. An emergency in ERCOT must be resolved with the resources physically located within ERCOT's boundaries, supplemented only by the trickle of power available through the DC ties.

This trade-off was understood from the beginning. It was accepted as the price of regulatory independence. For decades, it appeared to be a manageable risk. Texas had ample generation capacity, strong load growth attracted new investment, and weather emergencies, while occasionally stressful, did not approach the scale that would test the limits of the island strategy.

And then came Uri.

* * *

III. Winter Storm Uri: The Stress Test

The Event

The week of February 14-19, 2021, brought the most severe winter weather to the state of Texas in at least a generation, and arguably in modern recorded history. An Arctic air mass, driven south by a weakened polar vortex, pushed temperatures across the state to levels far below historical norms. Dallas recorded a low of minus two degrees Fahrenheit. Houston, a subtropical city, experienced temperatures in the teens. Even the Rio Grande Valley, along the Mexican border, saw freezing conditions.

The immediate consequence was an explosion of electricity demand. Texas homes are overwhelmingly heated with electricity — heat pumps and resistance heating — rather than natural gas furnaces. As temperatures plunged, electric heating loads surged to levels that the system had never been designed to accommodate. Winter peak demand in ERCOT had historically been well below summer peak demand, and the system was planned and built accordingly. The demand spike driven by Uri

exceeded even the most extreme winter planning scenarios.

But the demand spike was only half the crisis. The supply side collapsed simultaneously, in a cascading failure that ultimately proved far more consequential.

The Cascading Supply Failure

The physical failures during Uri were pervasive and cut across every generation technology.

Natural gas, which provided the largest share of ERCOT's generation capacity, suffered the most severe losses. Gas wells across West Texas and the Permian Basin froze. Gathering lines, which transport gas from wellheads to processing plants, froze. Processing plants, which remove impurities from raw gas before it can be burned in power plants, froze. Compressor stations on the interstate pipeline network, which maintain the pressure needed to transport gas over long distances, lost power — in many cases because the electricity they depended on was itself being curtailed. The result was a vicious feedback loop: power plants could not generate because they lacked gas, and gas facilities could not operate because they lacked power. Natural gas production in Texas declined by an estimated 50 percent or more during the worst of the storm.

Wind generation, which had grown to represent a significant share of ERCOT's installed capacity, also experienced major losses. Wind turbines across West Texas and the Panhandle froze, their blades coated in ice, their control systems and lubrication systems inoperable in the extreme cold. Wind output dropped precipitously at the very moment it was most needed. It should be noted, however, that wind generation in Texas is typically expected to contribute less during winter than during other seasons, and ERCOT's capacity planning did not rely on high winter wind output. The wind losses were significant in absolute terms but were dwarfed by the losses in natural gas generation.

Coal plants, which represented a smaller but still meaningful share of the generation fleet, suffered their own failures. Coal piles froze solid, making it impossible to feed coal into boilers through normal conveyor systems. Coal-handling equipment seized in the cold. Several coal units tripped offline due to frozen instrumentation and control systems.

Nuclear generation, often regarded as the most reliable baseload technology, was not immune. One of the two units at the South Texas Project Nuclear Generating Station tripped offline due to a frozen sensing line — a pressure-sensing instrument that, when it froze and sent erroneous readings, triggered an automatic reactor shutdown. The loss of over 1,300 megawatts of nuclear capacity at the height of the crisis compounded an already desperate situation.

At the nadir of the event, ERCOT lost approximately 52,000 megawatts of generation capacity — more than half of its installed fleet. Available generation fell below 45,000 megawatts against a demand that, even with emergency appeals for conservation, exceeded 69,000 megawatts.

The Near-Collapse

The scale of the supply-demand imbalance pushed ERCOT to the edge of a catastrophe far worse than

the blackouts that actually occurred. On the morning of February 15, grid frequency — the fundamental indicator of supply-demand balance in an alternating-current system, normally maintained at precisely 60 hertz — began to drop. When frequency falls, it means that demand is exceeding supply. Small frequency deviations are routine and are corrected by automatic generator controls. Large deviations are dangerous. If frequency falls too far, generators begin to trip offline automatically to protect themselves from damage, which further reduces supply, which drives frequency lower still — a cascading collapse that can, in the worst case, result in a complete blackout of the entire interconnection.

ERCOT's frequency dropped to 59.302 hertz, perilously close to the threshold at which automatic under-frequency load shedding relays would have disconnected large blocks of load and generation in an uncontrolled fashion. System operators later stated that the grid was approximately four minutes and thirty-seven seconds from a total collapse — a complete blackout that could have taken weeks or even months to fully restore through the painstaking process of black-start recovery.

To prevent that outcome, ERCOT ordered controlled load shedding — rolling blackouts — beginning on February 15. But the term "rolling" proved bitterly misleading. Because the supply deficit was so large and so persistent, the blackouts could not be rotated in the short intervals (typically 15-45 minutes) that the term implies. Millions of Texans lost power for periods ranging from hours to more than four consecutive days. In many cases, the same neighborhoods bore the burden of extended outages because the circuits serving them lacked critical infrastructure designations that would have protected them from curtailment.

The Human Cost

The human consequences of Winter Storm Uri were devastating. An estimated 246 people died, according to the Texas Department of State Health Services, though independent analyses have suggested the true toll may have been significantly higher — perhaps 700 or more — when indirect deaths from carbon monoxide poisoning, hypothermia, exacerbation of chronic medical conditions, and delayed medical care are included.

Millions of people endured days without heat in homes that had reached interior temperatures in the thirties and forties. Water systems across the state failed as pipes froze and burst, leaving communities without potable water for days or weeks after power was restored. Hospitals operated on backup generators while managing surges of hypothermia and carbon monoxide poisoning patients. The elderly, the medically vulnerable, and low-income communities — particularly those in poorly insulated housing — bore a disproportionate share of the suffering.

The economic costs were staggering. Total damages have been estimated at $80 billion to $130 billion when accounting for property damage, lost economic activity, emergency response costs, and the extraordinary prices charged in the wholesale electricity and natural gas markets during the crisis. Wholesale electricity prices in ERCOT were administratively set at the $9,000/MWh cap for approximately four consecutive days — a duration that the energy-only market design had never contemplated. Natural gas spot prices spiked to levels exceeding $400 per MMBtu, roughly 100 times

normal levels.

The financial consequences cascated through the electricity supply chain. Brazos Electric Power Cooperative, the largest generation and transmission cooperative in Texas, filed for bankruptcy. Griddy, a retail electricity provider that passed wholesale prices directly through to consumers, collapsed after its customers received electricity bills of $10,000 or more for a single week. The financial fallout consumed years of litigation and regulatory proceedings.

What Uri Revealed

Winter Storm Uri was not merely an operational failure. It was a systemic revelation — a stress test that exposed vulnerabilities embedded in the very architecture of the Texas electricity experiment.

First, and most fundamentally, Uri demonstrated the physical consequences of ERCOT's island strategy. At the peak of the crisis, ERCOT was importing the maximum available power through its DC ties — a few hundred megawatts. In a fully interconnected system, the massive generation reserves available in the Eastern Interconnection could have been marshaled to assist. Thousands of megawatts of emergency imports might have reduced or eliminated the need for load shedding. Isolation, the cornerstone of Texas's jurisdictional independence, was also a wall that kept help out.

Second, Uri exposed the limitations of the energy-only market design under conditions of extreme and prolonged scarcity. The ORDC and the offer cap were designed to produce high prices during brief periods of peak demand — the hottest hours of the hottest summer days. They were not designed for a scenario in which the supply deficit persisted for four or more days. The price signals worked, in the sense that prices reached and remained at the cap. But they worked too late to matter. No amount of price signaling can conjure a gas well back from a frozen state, defrost a wind turbine in a matter of hours, or restart a tripped nuclear unit during a winter storm. The scarcity pricing mechanism is an investment signal — it is supposed to work over years and decades, incentivizing the construction of resources that will be available when needed. It is not a real-time operational tool for managing a crisis that has already arrived.

Third, Uri revealed the dangerous interdependency between the electricity system and the natural gas supply chain — an interdependency that ERCOT's market design had not adequately addressed. The electricity-gas nexus is a feedback loop: gas plants need electricity to compress and transport gas, and the electricity system needs gas to generate power. When both systems are stressed simultaneously, the potential for cascading failure is enormous. ERCOT's market design treated gas supply as exogenous — as a given — rather than as a variable that could itself fail catastrophically.

Fourth, and perhaps most uncomfortably for the architects of the Texas experiment, Uri raised the question of whether the energy-only market had actually delivered the investment in winterized, reliable generation capacity that the system required. The generators that failed during Uri had, in many cases, not been winterized because the cost of winterization exceeded the expected revenue from winter scarcity events. This was, in a narrow sense, a rational economic calculation within the incentive structure of the energy-only market. But it was a calculation that proved catastrophically wrong when the

tail risk materialized.

* * *

IV. The Political and Legislative Aftermath

Accountability and Governance

The political fallout from Winter Storm Uri was swift and severe. The chair of the PUCT resigned. The remaining commissioners were replaced. The CEO of ERCOT was terminated. The ERCOT board of directors, which included several members who did not reside in Texas — a fact that generated intense public anger — was reconstituted through emergency legislation.

Governor Greg Abbott, who had initially deflected blame toward renewable energy (a characterization that was factually incomplete and politically convenient), eventually signed into law a sweeping set of reforms. The Texas Legislature, in its 2021 regular session, enacted Senate Bill 3 and a package of companion legislation that represented the most significant restructuring of Texas electricity policy since deregulation itself.

Senate Bill 3 and Weatherization

Senate Bill 3 addressed the most immediate and visible failure revealed by Uri: the lack of weatherization requirements for generation and gas supply infrastructure. The legislation directed the PUCT and the Railroad Commission of Texas (which regulates the oil and gas industry) to establish mandatory weatherization standards for electricity generators, transmission and distribution utilities, and natural gas facilities designated as critical infrastructure.

The implementation of these requirements proved contentious and complex. Defining which gas facilities qualified as "critical infrastructure" required mapping the interdependencies between the electricity and gas systems — an exercise that had never been systematically undertaken. Establishing weatherization standards required determining the appropriate planning temperature — the extreme cold scenario against which infrastructure should be hardened — a question with significant cost implications. Enforcement mechanisms had to be designed and staffed.

By subsequent winters, significant progress had been made. Generation facilities invested in insulation, heat tracing, wind breaks, and other cold-weather protections. The electric grid performed markedly better during Winter Storm Elliott in December 2022 and subsequent cold weather events, though critics noted that none of these events approached the severity of Uri.

The Performance Credit Mechanism: A Quasi-Capacity Construct

Perhaps the most consequential policy response to Uri was not SB 3's weatherization requirements but rather the PUCT's subsequent development of a Performance Credit Mechanism, or PCM — a market construct that represented a significant philosophical departure from the pure energy-only design that had defined ERCOT for two decades.

The PCM, adopted by the PUCT after extensive deliberation, creates a system in which load-serving entities are required to procure performance credits from generators that demonstrate the ability to perform during periods of system-wide scarcity. The mechanism is designed to provide an additional revenue stream to dispatchable generation resources — resources that can be relied upon to produce electricity when the system needs it most — beyond what they earn in the energy market alone.

The PCM is carefully constructed to avoid the label of a "capacity market." Texas policymakers, ideologically committed to the energy-only framework, have resisted framing the PCM as a capacity payment. They describe it instead as a performance-based reliability mechanism that rewards actual delivery of electricity during scarcity conditions rather than mere availability. The distinction is more than semantic: a traditional capacity market pays generators for being available to run; the PCM pays generators based on their demonstrated performance during designated scarcity hours.

Whether the PCM is, in substance, a capacity market by another name is a question on which reasonable analysts disagree. Its supporters argue that it preserves the essential character of the energy-only market while addressing the specific investment gap that Uri exposed — the underinvestment in reliable, weatherized, dispatchable capacity. Its critics, from both sides, offer competing objections. Free-market purists argue that the PCM is an unnecessary intervention that will distort the price signals of the energy market and begin the slide toward central planning. Consumer advocates and reliability hawks argue that the PCM does not go far enough — that a full capacity market, with mandatory reserve margins and enforceable performance standards, is the only mechanism that can ensure the level of reliability that a modern society requires.

The PCM's ultimate effectiveness will be judged over the coming decade as the mechanism matures, as its incentive effects on investment decisions become visible, and as the Texas grid faces future weather extremes. It represents, in either case, a significant concession: an acknowledgment by the state that adopted the purest energy-only market in the United States that energy-only pricing, even with the ORDC, may not be sufficient to ensure reliability.

* * *

V. ERCOT in Comparative Perspective

The Texas experiment, viewed in the context of American electricity policy more broadly, offers several

lessons that extend well beyond the state's borders.

The first is that market design is not merely an economic abstraction. It has physical consequences. The choice between an energy-only market and a capacity market, the level of the offer cap, the design of scarcity pricing mechanisms — these are decisions that determine, in very concrete terms, whether generators are built, whether they are winterized, and whether they are available when a winter storm arrives. The Texas experience demonstrates that the gap between theoretical elegance and operational reality can be measured in human lives.

The second lesson concerns the limits of isolation in an interconnected world. The Texas grid was designed to be self-sufficient, and for most of its history it has been. But the physics of extreme weather events do not respect jurisdictional boundaries, and the option value of interconnection — the ability to draw on distant resources during a local emergency — is substantial. Texas has purchased its regulatory independence at the cost of that option value, and Uri demonstrated that the cost can be very high.

The third lesson is about the relationship between markets and governance. The Texas experiment was, in part, a test of the proposition that electricity — a commodity with unique physical characteristics, including the inability to be economically stored at scale and the requirement for instantaneous balancing of supply and demand — could be governed primarily through market mechanisms rather than regulatory mandates. Uri did not disprove this proposition entirely, but it demonstrated that the boundary between market design and public safety regulation cannot be drawn as cleanly as the architects of the Texas experiment had hoped. Weatherization standards, performance requirements, and supply-chain resilience are not market outcomes; they are regulatory choices that markets alone may not produce.

Texas, for its part, has not abandoned the experiment. It has modified it — adding weatherization mandates, introducing the PCM, restructuring governance — while preserving the core elements of its energy-only market and its jurisdictional independence. Whether these modifications are sufficient to prevent a recurrence of Uri, or whether they represent way stations on a longer journey toward a more conventional market design, remains to be seen.

* * *

Conclusion

ERCOT stands as the boldest experiment in electricity market design in the United States, and perhaps in the developed world. It is a market built on principles of regulatory autonomy, competitive pricing, and minimal government intervention — principles that resonate deeply with the political culture of the state that created it.

The experiment has produced genuine achievements. Texas has attracted enormous investment in generation capacity, including an unmatched buildout of wind and solar resources. Wholesale electricity prices have been, on average, competitive with or lower than those in other deregulated markets. The

market has demonstrated remarkable flexibility in adapting to changing resource mixes and demand patterns.

But the experiment has also produced a catastrophic failure — one that killed hundreds of people, inflicted tens of billions of dollars in damages, and forced a fundamental reconsideration of the assumptions on which the market was built. The question that Winter Storm Uri posed to Texas, and that Texas has not yet fully answered, is whether an electricity system serving 27 million people can be reliably operated as an island, governed by scarcity pricing, and largely insulated from federal oversight — or whether the physical realities of the electric grid eventually demand a degree of interconnection, regulation, and centralized reliability assurance that the Texas experiment was designed to avoid.

The answer to that question will be written not in academic texts but in the performance of the Texas grid during the next extreme weather event. When it comes — and in a state prone to both brutal summers and, as climate patterns shift, increasingly severe winter storms, it will come — ERCOT's modified market design will face its next test. The residents of Texas, who have no choice but to depend on the grid their state has built, will be the ones who live with the result.

* * *

Key Concepts: energy-only market, capacity market, Operating Reserve Demand Curve (ORDC), system-wide offer cap, Value of Lost Load (VOLL), Peaker Net Margin, Cost of New Entry (CONE), scarcity pricing, merit order effect, jurisdictional isolation, Federal Power Act, interstate commerce, Performance Credit Mechanism (PCM), weatherization, load shedding, grid frequency, black-start recovery, electricity-gas interdependency.

Chapter 8: The Western Frontier and the EIM

Introduction: A Region Like No Other

The American West defies the organizational logic that governs electricity systems east of the Rocky Mountains. Where the Eastern Interconnection has consolidated into large regional transmission organizations — PJM stretching from the mid-Atlantic to Chicago, MISO spanning the Great Plains, SPP covering the southern wind belt — the Western Interconnection remains a mosaic of thirty-eight distinct balancing authorities, each managing its own slice of a grid that stretches from the Canadian border to Baja California, from the Pacific Coast to the front range of the Rockies. This fragmentation is not an accident of history but a reflection of the West itself: a region defined by vast distances, extraordinary geographic diversity, and a deeply rooted tradition of political independence.

The physical geography of the West created an electricity system unlike any other. The great federal dam-building projects of the twentieth century — Grand Coulee, Hoover, Glen Canyon, Bonneville — established massive publicly owned generation resources operated by federal power marketing administrations. The Bonneville Power Administration in the Pacific Northwest and the Western Area Power Administration across the interior West became anchors of a system in which public power played a role far larger than anywhere else in the country. Around these federal anchors grew a diverse ecosystem of investor-owned utilities, municipal systems, rural electric cooperatives, tribal utilities, and state-chartered authorities, each jealously guarding its autonomy and its preferential access to low-cost hydroelectric power.

Into this fragmented institutional landscape has come an energy transition of staggering speed and scale. California, the largest economy in the West and the fifth largest in the world, has pursued renewable energy deployment with an ambition unmatched by any other American state. The consequences of that pursuit — technical, economic, and political — have radiated outward across the entire Western Interconnection, reshaping relationships between states and utilities that have operated independently for decades. The "Duck Curve," a term that has entered the lexicon of every grid operator

in North America, is a California creation, born of the collision between aggressive clean energy policy and the operational realities of running a reliable power system. And the Energy Imbalance Market, the most significant institutional innovation in western electricity governance in a generation, is an attempt to manage the consequences of that collision without requiring the kind of wholesale institutional restructuring that the West has repeatedly rejected.

This chapter examines both phenomena in detail. It begins with the California Independent System Operator and the operational challenges created by the state's solar revolution, then turns to the incremental, trust-building process through which the broader West has begun to knit itself into a more integrated market — without surrendering the independence that defines it.

* * *

8.1 CAISO: The Single-State ISO

8.1.1 Origins and Governance

The California Independent System Operator occupies a unique position in American electricity governance. It is the only independent system operator in the United States that operates within the boundaries of a single state, and its governance structure reflects that singularity in ways that have profound consequences for the broader West.

CAISO was created in 1998 as part of California's electricity restructuring legislation, Assembly Bill 1890, which deregulated the state's electricity market. Unlike other ISOs and RTOs, whose boards are selected through FERC-approved stakeholder processes designed to balance the interests of generators, transmission owners, load-serving entities, and consumers, CAISO's board of governors is appointed by the Governor of California and confirmed by the state Senate. This means that CAISO's leadership is ultimately accountable to California's elected officials and, by extension, to California's policy priorities — including the state's ambitious climate and clean energy goals.

This governance structure is simultaneously CAISO's greatest asset within California and its greatest liability beyond the state's borders. Within California, it ensures that the grid operator's priorities are aligned with state energy policy, enabling the kind of coordinated planning that has facilitated the rapid deployment of renewable energy. But for utilities and regulators in other western states, the prospect of participating in a market governed by an entity answerable to California's legislature raises fundamental questions of sovereignty and representation. Why, officials in Arizona or Wyoming or Oregon ask, should their ratepayers be subject to market rules shaped by California's political priorities? This governance question, as we shall see, has become the central obstacle to full market integration in the West.

CAISO manages approximately 80 percent of California's electricity load, operating the transmission grid for the state's three large investor-owned utilities — Pacific Gas and Electric, Southern California Edison, and San Diego Gas and Electric. It runs both day-ahead and real-time energy markets, manages congestion on the transmission system, and oversees resource adequacy requirements intended to ensure that sufficient generation capacity is available to meet demand under a range of conditions. The territory it manages is large by the standards of a single state — roughly 26,000 circuit miles of transmission — but small by the standards of other ISOs. PJM's footprint is more than five times larger; MISO's is nearly four times. This relative compactness, combined with California's position at the western edge of the continent with limited transmission ties to neighboring systems, has made CAISO's operational challenges particularly acute as the state's resource mix has undergone a dramatic transformation.

8.1.2 The Renewable Revolution and the Birth of the Duck Curve

California's renewable energy ambitions are codified in a series of increasingly aggressive renewable portfolio standards. The state first established an RPS in 2002, requiring utilities to procure 20 percent of their electricity from renewable sources by 2017. Subsequent legislation accelerated the timeline and raised the target: 33 percent by 2020, 50 percent by 2030 under Senate Bill 350 in 2015, 60 percent by 2030 under Senate Bill 100 in 2018, with an ultimate goal of 100 percent clean electricity by 2045. These mandates, combined with the Investment Tax Credit at the federal level and California's own incentive programs, unleashed a wave of solar photovoltaic deployment that fundamentally altered the shape of the state's electricity demand curve.

To understand the Duck Curve, one must first understand the concept of net load. Gross load is the total electricity demand on the system at any given moment — the sum of all the air conditioners, factories, lights, appliances, and electric vehicle chargers drawing power. Net load is gross load minus the output of variable renewable resources — primarily solar and wind. Net load represents the demand that must be met by dispatchable resources: natural gas plants, hydroelectric facilities, nuclear stations, imports from neighboring systems, and, increasingly, battery storage. It is net load, not gross load, that grid operators must manage in real time.

Before the solar revolution, California's net load curve on a typical spring day looked roughly like a gentle hill: demand rising in the morning, peaking in the late afternoon, and declining through the evening. By the mid-2010s, massive solar deployment had carved a deep valley into the middle of that hill. On sunny days, solar output surged beginning around 9:00 a.m., depressing net load to levels far below what the system had been designed to serve. By early afternoon, net load could fall to a fraction of its morning level — and on mild spring days with abundant sunshine, it could approach zero or even turn negative, meaning that renewable generation was producing more electricity than California could consume.

Then, beginning around 4:00 p.m. as the sun descended toward the Pacific, solar output dropped precipitously. Over the course of roughly three hours, the system lost tens of thousands of megawatts of

solar generation just as residential demand was climbing — people returning home, turning on air conditioning, cooking dinner, charging vehicles. The net load curve shot upward in a steep ramp that required the system to bring enormous quantities of dispatchable generation online in a very short period.

When CAISO first published projections of this phenomenon in 2013, plotting the net load curves for successive years on a single chart, the resulting shape — a fat belly in the middle of the day, a long neck rising steeply in the late afternoon — bore an unmistakable resemblance to a duck. The "Duck Curve" became one of the most widely reproduced graphics in energy policy, a simple visual that captured the essential operational challenge of high solar penetration.

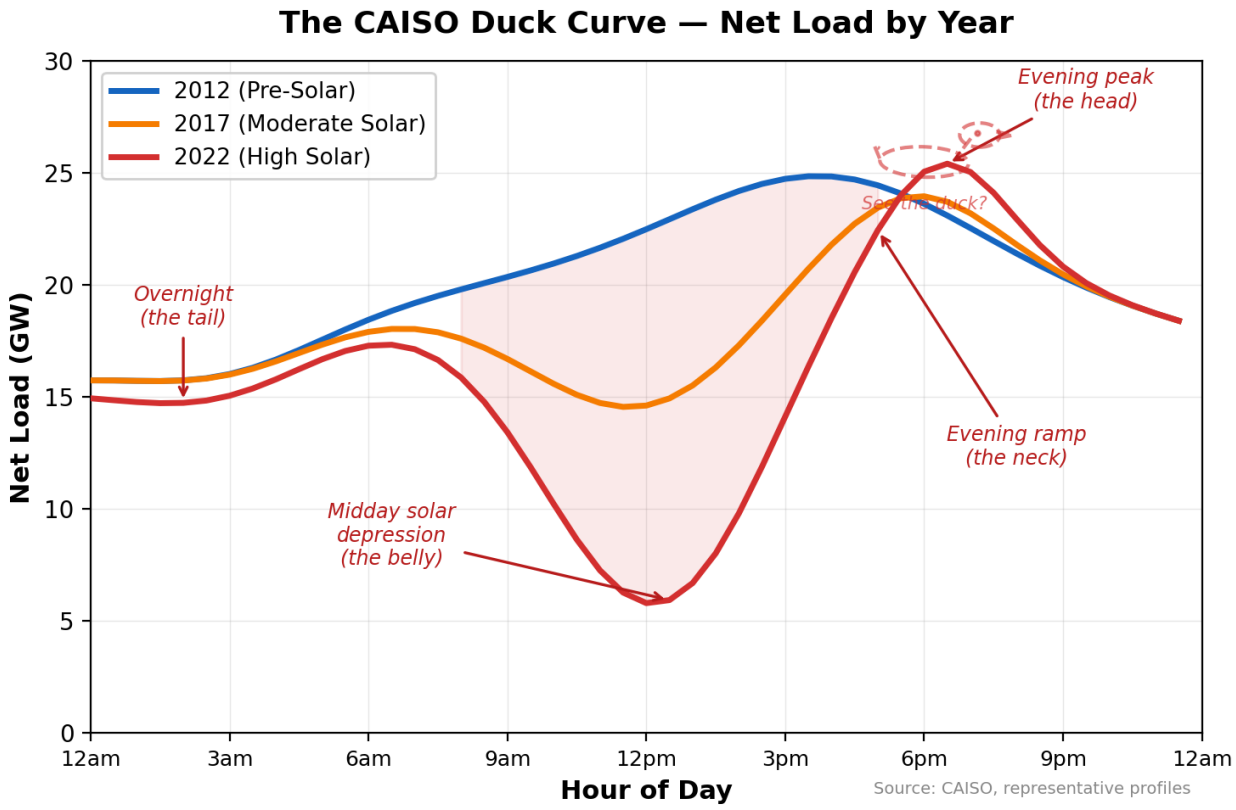

Figure 8.2: The CAISO Duck Curve — Net Load Profiles Under Increasing Solar Penetration (Source: CAISO)

What was a projection in 2013 has become an increasingly extreme reality. By the early 2020s, the duck had grown substantially fatter. Midday net load troughs on spring days routinely fell below 5,000 megawatts in a system with a summer peak capacity need exceeding 45,000 megawatts. On some days, net load briefly went negative — a condition that meant California had more renewable generation available than it had demand to absorb, even after exporting as much power as its transmission ties to neighboring states would allow. The evening ramp steepened further: CAISO operators faced the challenge of ramping dispatchable generation upward by 15,000 to 20,000 megawatts or more in a three-hour window, a rate of change for which the system's legacy gas fleet was not originally designed.

8.1.3 Operational Consequences

The Duck Curve created a cascade of operational challenges, each compounding the others.

Curtailment. When net load approaches zero or turns negative, the system must curtail — that is, deliberately reduce — renewable generation. Solar plants are instructed to reduce their output even though the sun is shining and the marginal cost of their electricity is essentially zero. Curtailment represents wasted clean energy and undermines the economics of renewable projects. In 2022 and subsequent years, California curtailed millions of megawatt-hours of solar and wind generation annually, a figure that has continued to grow as new solar capacity comes online faster than load growth and storage deployment can absorb it. Each curtailed megawatt-hour is a megawatt-hour that cannot displace fossil fuel generation, representing both an economic loss and a missed opportunity for emissions reduction.

Negative wholesale prices. In a competitive wholesale market, an oversupply of zero-marginal-cost renewable energy pushes clearing prices downward. When supply exceeds demand and curtailment has not yet fully resolved the imbalance, prices can turn negative — generators are effectively paying the market to take their electricity. Negative prices are not merely a curiosity; they distort investment signals, erode the revenue streams of all generators (including the renewables whose output created the condition), and create perverse incentives. A gas plant that must stay online to provide evening ramping capability may lose money during the midday hours when it is needed least, raising questions about whether such plants can remain economically viable without capacity payments or other out-of-market revenue.

Ramping requirements. The steep evening ramp places a premium on resources that can increase output quickly and reliably. Simple-cycle gas turbines, which can start from cold and reach full output in ten to fifteen minutes, become enormously valuable during the ramp window, even if they sit idle for much of the day. Older steam turbines and combined-cycle plants, which ramp more slowly, may struggle to meet the rate of change required. The evening ramp has also forced CAISO to rethink how it schedules resources in the hours leading up to sunset, pre-positioning generation in advance of the ramp to avoid reliability emergencies.

The growing importance of battery storage. The Duck Curve has created an almost textbook economic case for battery energy storage. Batteries can charge during the midday period when solar output is abundant and prices are low (or negative), absorbing excess generation that would otherwise be curtailed. They can then discharge during the evening ramp, providing fast-responding capacity precisely when the system needs it most. The economics are compelling: buy low, sell high, on a daily cycle driven by an entirely predictable pattern of solar generation and demand.

California has responded with a storage deployment push that has no precedent in the industry. Following a 2013 mandate requiring utilities to procure 1,325 megawatts of storage, the state's storage portfolio has grown exponentially. By the mid-2020s, California had deployed well over 10,000 megawatts of battery storage capacity, overwhelmingly four-hour lithium-ion systems. These batteries have become a critical component of the evening net load ramp, providing thousands of megawatts of discharge capacity during the critical 4:00 to 9:00 p.m. window. On some evenings, batteries have

provided more than 5,000 megawatts of output simultaneously, rivaling the contribution of a large nuclear plant.

Yet storage is not a panacea. Four-hour batteries, as their name implies, can sustain full output for only four hours. On the longest summer evenings, when demand remains elevated well into the night, batteries that began discharging at 4:00 p.m. may be depleted by 8:00 p.m., leaving the system short of capacity during the final hours of the peak. Longer-duration storage — six, eight, or twelve hours — is technically feasible but significantly more expensive per unit of capacity. The question of optimal storage duration remains one of the most active areas of resource planning analysis in California and across the industry.

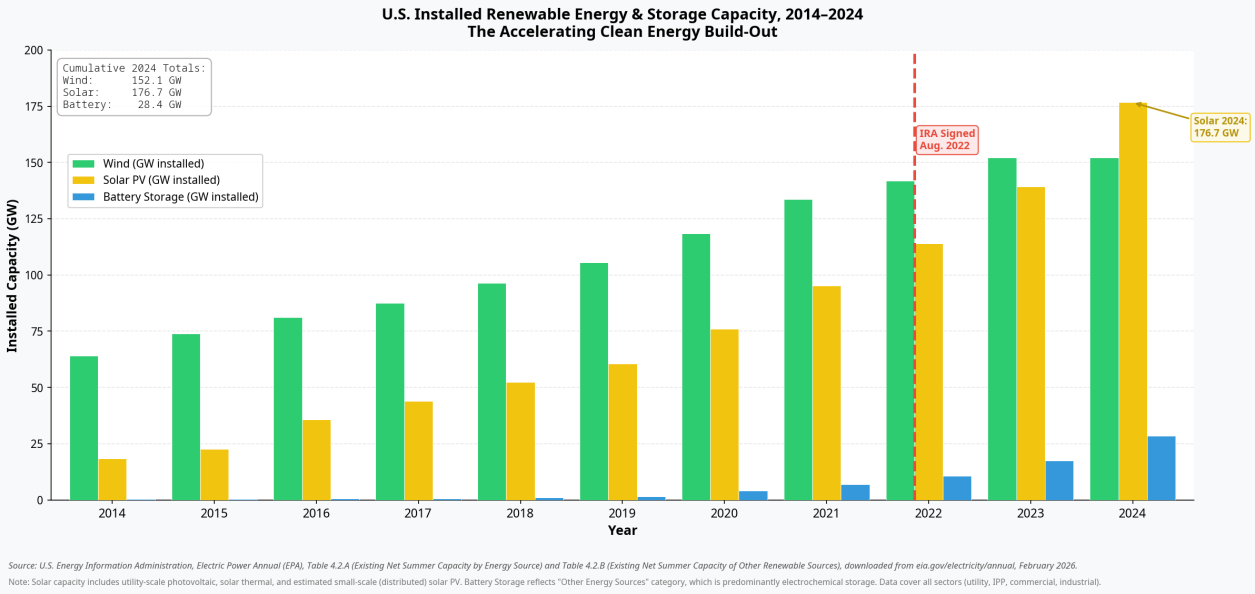

Figure 8.1: U.S. Installed Renewable Energy and Storage Capacity, 2014–2024 (Source: EIA Electric Power Annual)

8.1.4 The August 2020 Rolling Blackouts: A Case Study in Resource Adequacy Failure

On August 14 and 15, 2020, CAISO ordered the first rolling blackouts in California since the energy crisis of 2000–2001. Approximately 800,000 customers lost power in rotating outages lasting one to two and a half hours. The immediate cause was an extreme heat wave that settled over the entire western United States, driving air conditioning loads to near-record levels across the region. But the root cause, as a subsequent joint investigation by CAISO, the California Public Utilities Commission, and the California Energy Commission concluded, was a failure of resource adequacy planning to keep pace with the transformation of the state's resource mix.

Several factors converged. First, resource adequacy planning had not fully accounted for the risk of region-wide extreme heat events, which simultaneously increased demand across the entire West and

reduced the availability of imports from neighboring systems that were themselves under stress. California had long relied on imports to meet a significant portion of its peak demand, but when every system in the region was experiencing high loads, surplus power for export simply did not exist.

Second, the retirement of once-through-cooling gas plants along the California coast — retirements driven by legitimate environmental concerns about their water intake systems — had removed dispatchable capacity from the fleet without sufficient replacement. The new solar resources that had come online in their stead produced abundantly during the day but were of no help during the evening peak when the blackouts occurred.

Third, the resource adequacy framework had used planning assumptions that underestimated the risk of coincident high loads and low renewable output during extreme heat events. The planning reserve margin, the buffer of capacity above expected peak demand intended to account for contingencies, proved insufficient for the conditions actually encountered.

The August 2020 blackouts became a powerful cautionary tale for the industry. They demonstrated that high renewable penetration, while essential for decarbonization, requires careful attention to the adequacy and availability of dispatchable resources — including storage, demand response, and gas generation — during the hours when renewable output is low and demand is high. California responded with emergency procurement of additional capacity, reforms to its resource adequacy framework, and an accelerated push for battery storage that has shaped the state's resource mix in the years since.

The episode also carried political weight. Twenty years after the energy crisis that had derailed deregulation and destroyed political careers, rolling blackouts remained the third rail of California energy politics. The 2020 events intensified political pressure on CAISO and state regulators to ensure reliability even as they pursued increasingly ambitious clean energy targets, a tension that continues to shape policy debates.

* * *

8.2 The Energy Imbalance Market: Integration Without Consolidation

8.2.1 Why the West Resisted RTOs

To understand the significance of the Energy Imbalance Market, one must first understand why the Western Interconnection has been so resistant to the kind of wholesale market integration that has transformed electricity systems east of the Rockies.

The defining event is the California Energy Crisis of 2000–2001. In the summer of 2000 and

through the spring of 2001, wholesale electricity prices in California skyrocketed. Rolling blackouts struck the state. Pacific Gas and Electric filed for bankruptcy. The state's largest utility, Southern California Edison, teetered on the edge of insolvency. The state government, under Governor Gray Davis, spent billions of dollars in emergency power purchases, eventually contributing to a fiscal crisis that played a role in Davis's recall from office in 2003.

The causes of the crisis were complex — a confluence of flawed market design, a drought that reduced hydroelectric output across the West, inadequate generation capacity, and, critically, deliberate market manipulation by energy trading companies, most notoriously Enron. Traders engaged in schemes with colorful names — "Death Star," "Fat Boy," "Get Shorty" — that artificially created congestion, withheld supply, and drove prices to extraordinary levels. The eventual exposure of these schemes, the criminal prosecution of several Enron traders, and the broader collapse of Enron itself became defining events in American energy policy.

The political fallout was devastating and enduring. Within California, the crisis discredited electricity deregulation and created deep skepticism of wholesale electricity markets that persists to this day. Outside California, the crisis created an equally powerful narrative: that California's reckless experiment with deregulation had destabilized the entire western grid, and that closer integration with California's market would expose other states' ratepayers to the same risks. The crisis poisoned the well for regional market integration for a generation.

Beyond the crisis, structural features of the Western Interconnection militate against consolidation. The diversity of utility ownership structures is far greater than in the East. The Bonneville Power Administration, a federal agency within the Department of Energy, markets the output of thirty-one federal dams in the Columbia River Basin and operates 15,000 miles of high-voltage transmission. It provides roughly a third of the Pacific Northwest's electricity, nearly all of it carbon-free hydropower, at rates well below those of neighboring investor-owned utilities. BPA's customers — the publicly owned utilities and cooperatives of the Northwest — have fought fiercely to protect their access to low-cost federal hydropower and have viewed RTO formation as a potential threat to their preferential rates. The Western Area Power Administration plays a similar role across the interior West, marketing power from major federal projects in the Colorado River Basin and the Missouri River Basin.

Western states also exhibit a fierce tradition of regulatory sovereignty. State public utility commissions in the West have historically exercised tight control over resource planning, procurement, and rates within their jurisdictions. Joining an RTO would require ceding some of that control to a regional entity, a prospect that state regulators and legislatures have viewed with suspicion. The political cultures of states like Wyoming, Montana, Idaho, and Utah — resource-rich, skeptical of federal authority, protective of state prerogatives — have reinforced this resistance.

The result, as the twenty-first century's second decade began, was a Western Interconnection that remained operationally balkanized. Thirty-eight balancing authorities managed their own systems, scheduling bilateral power transactions across a web of transmission contracts. The inefficiencies were substantial: each balancing authority maintained its own reserves, scheduled its own generation, and managed its own imbalances, resulting in duplicated costs and suboptimal dispatch. Power that was cheap and abundant in one corner of the West might go unused while a neighboring system ran

expensive gas plants, simply because no market mechanism existed to facilitate the transfer.

8.2.2 The EIM: An Incremental Solution

The Energy Imbalance Market, launched in November 2014, was designed to capture some of the benefits of regional market integration while avoiding the institutional upheaval of full RTO formation. Its genius lies in its modesty: the EIM is not a full wholesale market. It does not conduct a day-ahead energy auction. It does not manage long-term transmission rights. It does not oversee resource adequacy requirements. It does not require participating utilities to turn over operational control of their transmission systems. It does only one thing, but it does it very well: it optimizes real-time energy dispatch across participating balancing authorities.

The mechanics work as follows. Each participating balancing authority continues to plan and schedule its own generation fleet on a day-ahead and hourly basis, just as it always has. But in real time — at five-minute and fifteen-minute intervals — the EIM's optimization algorithm looks across all participating territories and identifies opportunities to redispatch generation more efficiently. If a utility in the Pacific Northwest has excess low-cost hydroelectric output at the same moment that a utility in the Desert Southwest is running an expensive gas peaker, the EIM facilitates a transfer: the hydro output displaces the gas generation, and both parties benefit. The hydro producer earns a higher price than it would have received in its own system, and the gas-dependent utility pays a lower price than its peaker would have cost.

The EIM began with just two participants: CAISO, which operates the market platform, and PacifiCorp, the large investor-owned utility serving portions of six western states. PacifiCorp's participation was driven in part by its need to integrate growing quantities of wind energy in Wyoming and Oregon and in part by a pragmatic calculation that the EIM's benefits would outweigh its costs — a calculation that the first year of operation quickly validated.

The early success of the EIM generated a virtuous cycle of participation. Each new entrant expanded the geographic footprint of the market, creating more opportunities for beneficial trades and increasing the value of participation for existing members. NV Energy joined in 2015. Arizona Public Service and Puget Sound Energy followed in 2016. Portland General Electric, Idaho Power, and Powerex (the trading arm of BC Hydro) joined in subsequent years. By the early 2020s, the EIM encompassed the vast majority of load in the Western Interconnection, stretching from British Columbia to New Mexico, from the Pacific Coast to the eastern slopes of the Rockies.

8.2.3 Documented Benefits

The benefits of the EIM have been carefully tracked and publicly reported, a transparency that has been critical to building trust and encouraging further participation.

Cost savings. Cumulative benefits have exceeded three billion dollars since the market's inception, a figure that represents the difference between the actual cost of serving load under the EIM's optimized

dispatch and the cost that would have been incurred had each balancing authority dispatched independently. These savings accrue to ratepayers across the West, providing a tangible, quantifiable return on participation.

Reduced curtailment. The EIM has significantly reduced the curtailment of renewable energy in California. Before the EIM, excess solar generation in California had no market outlet; it was simply curtailed. The EIM provides a mechanism for that excess energy to flow to neighboring systems that can absorb it, displacing fossil fuel generation and reducing waste. On sunny spring days, the EIM routinely facilitates transfers of thousands of megawatts of solar energy from California to systems across the West.

Inter-regional transfers. The EIM has enabled a volume of inter-regional energy transfers that would have been difficult or impossible to arrange through traditional bilateral scheduling. The five-minute dispatch interval allows the market to respond to rapidly changing conditions — a sudden cloud shadow over a solar field, an unexpected generator trip, a shift in wind patterns — with a speed and precision that bilateral trading cannot match.

Greenhouse gas reductions. By enabling the substitution of low-cost renewable energy for fossil fuel generation across the West, the EIM has produced measurable reductions in carbon dioxide emissions. The market includes a specific mechanism for tracking and accounting for greenhouse gas emissions associated with energy imports into California, ensuring that the state's climate policies are respected even as power flows freely across borders.

Reliability enhancement. The EIM has improved reliability by enabling participating balancing authorities to share reserves and lean on one another during system stress events. When one system experiences an unexpected generator outage, the EIM can automatically redispatch generation across the broader footprint to fill the gap, reducing the need for each individual system to maintain as large a reserve margin.

8.2.4 Building Trust

Perhaps the most important benefit of the EIM has been institutional rather than economic: it has built trust. For entities that had spent two decades avoiding entanglement with California's electricity market, the EIM provided a low-risk proving ground. Participation is voluntary. Entities retain full control of their transmission systems and generation fleets. They can withdraw if the market does not serve their interests. The costs of participation — primarily the software and staffing needed to interface with the market platform — are modest relative to the benefits.

This trust-building function cannot be overstated. The Western Interconnection's resistance to integration was never primarily about economics; it was about politics, institutional culture, and the lingering trauma of the California Energy Crisis. The EIM has demonstrated, through years of successful operation, that a market platform operated by CAISO can deliver benefits to all participants without imposing California's policy preferences on others or creating the conditions for market manipulation. Each year of successful operation, each quarterly benefits report, each new entity joining the market, has

incrementally eroded the wall of resistance that the energy crisis erected.

* * *

8.3 The Next Frontier: Day-Ahead Markets and the Governance Question

8.3.1 From Real-Time to Day-Ahead: EDAM and Markets+

The success of the EIM inevitably raised a question: if real-time dispatch optimization produces billions of dollars in benefits, how much more value could be captured by extending that optimization to the day-ahead time frame?

Day-ahead markets are where the bulk of wholesale energy transactions occur. In existing ISOs and RTOs, the day-ahead market clears a supply-and-demand auction for each hour of the following day, establishing schedules and prices that guide generation commitment decisions. These markets capture efficiencies that real-time markets alone cannot: they enable more efficient unit commitment (deciding which power plants to start up, a process that involves significant lead times and start-up costs), more effective management of transmission congestion, and better integration of demand-side resources.

Two competing proposals have emerged to bring day-ahead market functionality to the West.

The Extended Day-Ahead Market (EDAM). Proposed by CAISO, EDAM would extend the existing day-ahead market run by the ISO to encompass voluntary participants across the West, mirroring the structure of the EIM. Participants would submit supply offers and demand bids into a centralized day-ahead market, which would produce optimized schedules and clearing prices for each hour of the following day. EDAM would build on the existing EIM platform and participation base, leveraging the relationships, trust, and technical infrastructure that a decade of EIM operation has established.

Markets+. Proposed by the Southwest Power Pool, Markets+ offers an alternative day-ahead market platform governed by an independent, FERC-jurisdictional entity rather than by CAISO. SPP, which operates a full RTO in the Eastern Interconnection, brings decades of experience running wholesale markets and a governance model that many western entities view as more inclusive and less California-centric than CAISO's.

The competition between EDAM and Markets+ has become one of the most consequential institutional debates in American electricity in recent years. The choice is not merely technical — both platforms could deliver significant efficiency gains — but fundamentally about governance, trust, and the political economy of regional integration.

8.3.2 The Governance Debate

The governance question sits at the heart of the western market integration debate, and it is worth examining in detail because it illustrates a broader tension in American federalism as applied to electricity.

CAISO's board, as noted, is appointed by the Governor of California. This means that the entity operating both the EIM and the proposed EDAM is ultimately governed by officials responsive to California's political priorities. Those priorities include some of the nation's most aggressive decarbonization targets. For many western entities, this creates an unacceptable risk: that market rules, operating procedures, and investment signals will be shaped by California's climate policy in ways that disadvantage states with different priorities.

The concern is not entirely hypothetical. California's greenhouse gas policies already influence EIM operations through the market's GHG tracking mechanism, which attributes emissions to energy imports into California and effectively imposes a carbon cost on those transactions. While this mechanism operates within the existing legal framework of California's cap-and-trade program, it illustrates how a single state's policies can ripple through a regional market.

Western stakeholders have proposed various governance reforms. The most prominent would create an independent, FERC-jurisdictional governing body for the western market — a board selected through a stakeholder process, similar to those governing PJM, MISO, or SPP, rather than appointed by any single state's governor. California legislation has been introduced at various points to facilitate this transition, but the politics are delicate: California legislators are reluctant to surrender control of an entity that has served the state's interests effectively, while non-California stakeholders view independent governance as a prerequisite for their participation in a day-ahead market.

SPP's Markets+ proposal gains its appeal precisely from this governance dynamic. SPP is already governed by a FERC-jurisdictional stakeholder board with representation from across its footprint. Western entities that join Markets+ would participate in a governance structure designed from the outset to balance diverse interests, rather than one retrofitted onto a single-state ISO. For entities like the Bonneville Power Administration, which serves customers across four states and operates under a unique federal mandate, the appeal of a governance structure not dominated by any single state is considerable.

The risk of the competing proposals is fragmentation. If some western entities join EDAM and others join Markets+, the West could end up with two overlapping day-ahead markets, each covering only a portion of the interconnection and neither capturing the full efficiency gains that a single integrated market would deliver. This outcome — sometimes called the "seams" problem — would require complex coordination protocols between the two markets and could leave significant value on the table.

8.3.3 What Full Integration Might Look Like

The endgame, for many analysts and stakeholders, is a single organized wholesale market covering the

entire Western Interconnection — a "Western RTO" in all but name, if not in formal legal structure. Such a market would bring the West into rough parity with the eastern regions, enabling fully optimized generation dispatch, coordinated transmission planning, unified reliability standards, and a single set of market rules governing competition among generators.

The benefits of full integration would be substantial. Studies have estimated that a fully integrated western market could produce annual savings of one to two billion dollars or more, primarily through more efficient use of the West's diverse resource base. The complementarity of western resources is striking: California solar peaks in midday; Wyoming wind blows strongest at night and in winter; Pacific Northwest hydro is most abundant in spring and early summer; Desert Southwest solar persists into the evening hours. A fully integrated market would enable these resources to serve load across the entire region based on real-time conditions, rather than the happenstance of bilateral contracts and historical transmission agreements.

Full integration would also enhance reliability. The August 2020 blackouts in California demonstrated the risk of a system that depends on imports but lacks a market mechanism to ensure their availability during system-wide stress events. A single western market would internalize those interdependencies, enabling coordinated resource adequacy planning and reserve sharing across the entire interconnection.

But full integration would also require compromises that have thus far proven politically impossible. It would require California to accept governance reforms that dilute its control over the market operator. It would require federal power marketing administrations to reconcile their statutory obligations with market participation rules. It would require state regulators across the West to accept a degree of shared sovereignty over resource planning and market oversight. And it would require all parties to move beyond the trauma of the California Energy Crisis and the institutional habits of independence that have defined western electricity governance for a century.

8.3.4 The Path Forward

The most likely near-term trajectory is continued incrementalism. The EIM has demonstrated that voluntary, step-by-step integration can produce meaningful benefits without requiring a grand institutional bargain. EDAM, if successfully implemented, would represent the next step on this path — a significant expansion of market functionality that still preserves the voluntary, balancing-authority-centered structure of the EIM.

The competition with Markets+ may ultimately prove productive rather than destructive. The existence of a credible alternative to CAISO governance has created pressure for governance reform that might not otherwise exist. California policymakers, faced with the prospect of losing potential EDAM participants to a competing platform, have shown greater willingness to entertain governance changes that would make CAISO more palatable to the broader West. The dynamic is analogous to market competition itself: the threat of losing market share can motivate reforms that incumbents would otherwise resist.

Whether the West ultimately consolidates into a single market, settles into a two-market structure, or finds some other arrangement remains genuinely uncertain. What is clear is that the status quo — thirty-eight independent balancing authorities managing their systems in isolation — is no longer tenable. The scale of renewable energy deployment across the West, the operational challenges it creates, and the efficiency gains available through coordination are simply too large to ignore. The Duck Curve, born in California, has spread across the West as solar deployment has expanded beyond California's borders. The need for the kind of flexible, responsive dispatch that markets provide has become universal.

* * *

Conclusion: The Western Paradox

The Western Interconnection presents a paradox. It is the region of the United States with the greatest natural potential for renewable energy — vast solar resources in the Desert Southwest, world-class wind resources on the Great Plains and in the Columbia River Gorge, enormous hydroelectric capacity in the Pacific Northwest, geothermal resources in the Great Basin. Realizing that potential fully requires the kind of regional coordination that the West has historically resisted. The Duck Curve is, in a sense, a symptom of this paradox: it exists because California has deployed solar at enormous scale within a system that lacks the geographic and institutional breadth to absorb it efficiently.

The EIM has begun to resolve this paradox, but only partially. It has demonstrated that western entities can cooperate through a market mechanism without surrendering their autonomy. It has produced billions of dollars in documented benefits. It has reduced renewable curtailment, improved reliability, and built a foundation of trust that did not exist a decade ago. But the EIM captures only a fraction of the efficiency gains available through full market integration, and its real-time-only structure leaves the larger prize — optimized day-ahead commitment, coordinated transmission planning, unified resource adequacy — on the table.

The coming years will determine whether the West can build on the EIM's success to achieve deeper integration, or whether the institutional and political obstacles that have defined the region for a generation will prove insurmountable. The stakes are high: the West's ability to achieve its clean energy goals, maintain reliability in the face of growing climate stress, and keep electricity affordable for the more than eighty million people who depend on the Western Interconnection may well hinge on the outcome of debates about market design and governance that, to the uninitiated, might seem arcane. They are anything but. They are debates about how a vast, diverse, fiercely independent region manages one of the most important infrastructure systems in modern civilization — and whether it can do so in a way that honors both the physics of the grid and the politics of federalism.

* * *

Key Concepts from This Chapter:

- **Net load** — Total electricity demand minus variable renewable generation; the demand that must be met by dispatchable resources.
- **Duck Curve** — The characteristic shape of the net load curve under high solar penetration, featuring a deep midday trough and steep evening ramp.
- **Curtailment** — The deliberate reduction of renewable energy output when supply exceeds demand and export capacity.
- **Energy Imbalance Market (EIM)** — A real-time energy market that optimizes dispatch across voluntary participants in the Western Interconnection at five- and fifteen-minute intervals.
- **Extended Day-Ahead Market (EDAM)** — A proposed expansion of CAISO's market to include day-ahead energy scheduling for voluntary western participants.
- **Markets+** — A competing day-ahead market proposal operated by the Southwest Power Pool with independent, FERC-jurisdictional governance.
- **Resource adequacy** — The planning framework ensuring sufficient generation capacity is available to meet demand under expected and stressed conditions.
- **Balancing authority** — An entity responsible for maintaining the real-time balance between electricity supply and demand within its territory.

* * *

Chapter 9: The Northeast and Midwest Corridors

Introduction: Two Archetypes of the American Grid Challenge

The four regional transmission organizations and independent system operators examined in this chapter — ISO New England (ISO-NE), the New York Independent System Operator (NYISO), the Midcontinent Independent System Operator (MISO), and the Southwest Power Pool (SPP) — collectively serve more than 170 million Americans across a geographic expanse stretching from the Maine coastline to the Montana plains, from the Canadian border to the Gulf of Mexico. Yet for all their shared commitment to reliability and competitive wholesale markets, these four organizations confront two fundamentally different archetypes of the grid challenge, archetypes that in many respects define the central tensions of the twenty-first-century American power system.

The first archetype is the problem of *dense urban consumption*. ISO-NE and NYISO operate in regions where electricity must be delivered into some of the most congested, space-constrained, and politically contested load pockets on the continent. In southern New England and downstate New York, aging infrastructure threads through corridors that were laid out a century ago, hemmed in by waterways, highways, and dense settlement patterns that make expansion excruciatingly difficult. Generation siting is a perpetual struggle. Transmission upgrades require navigating not merely engineering constraints but also layers of municipal, state, and federal permitting in communities with sophisticated, well-organized opposition to new infrastructure. The fuel supply chain itself — particularly the natural gas pipeline network — was not designed for the scale of demand that has emerged as gas-fired generation has displaced coal and nuclear. The result is a set of markets that must constantly manage scarcity: scarcity of transmission capacity, scarcity of generation sites, scarcity of fuel during peak conditions, and increasingly, scarcity of political and social license to build.

The second archetype is the problem of *remote rural generation*. MISO and SPP operate across the vast interior of the continent, where some of the finest wind and solar resources on Earth are located in places with minimal local electricity demand. The Great Plains, from the Texas Panhandle through

western Kansas and the Dakotas, experience capacity factors for wind generation that rival or exceed those of many conventional power plants. Yet the load centers — Chicago, Minneapolis, St. Louis, the industrial Midwest, the Gulf Coast petrochemical corridor — are hundreds of miles away, separated by state lines, utility service territories, and a transmission network that was designed for a fundamentally different pattern of generation and consumption. The challenge here is not squeezing power into a congested urban pocket but rather building the long-distance transmission backbone necessary to move vast quantities of renewable energy from where the wind blows to where the people live.

These two archetypes are not merely regional curiosities. They represent the twin poles of a national challenge: the American power system must simultaneously retrofit its densest urban cores for a decarbonized future and construct a new continental-scale transmission network to harvest its best renewable resources. The solutions developed in these four regions — in market design, transmission planning, capacity procurement, and regulatory innovation — will shape the trajectory of the broader energy transition. This chapter examines each in turn.

* * *

I. ISO-NE and NYISO: Managing Aging Infrastructure and High-Density Urban Load

The Northeastern Grid: A Portrait of Constraint

The six New England states and New York share a set of characteristics that make their electric grids among the most challenging to operate and reform in the United States. Population density is high, particularly in the corridor from Boston through New York City, and electricity demand is concentrated in urban and suburban areas where land is scarce and expensive. The transmission network, much of it dating to the mid-twentieth century, was designed around a now-vanishing fleet of large central-station generators — coal plants, oil-fired steam units, and nuclear stations — sited at coastal and riverine locations that provided cooling water and fuel access. As these plants retire, they leave behind not only capacity deficits but also the loss of critical grid services such as voltage support and frequency response that the transmission system was engineered to rely upon.

The region's growing dependence on natural gas has introduced a new and acute vulnerability. Unlike coal, which can be stockpiled on-site, or nuclear fuel, which is loaded in multi-year cycles, natural gas arrives through a pipeline network that serves both electric generation and the region's enormous space-heating demand. During extreme cold weather events — the so-called "bomb cyclones" and polar vortex intrusions that periodically descend on the Northeast — gas demand for heating spikes at precisely the moment when electricity demand also surges, creating a direct competition for pipeline

capacity. The region's limited liquefied natural gas import infrastructure provides a partial buffer, but at significantly higher cost. The resulting fuel-security challenge has become a defining preoccupation of both ISO-NE and NYISO.

Layered atop these physical constraints are some of the most ambitious state-level clean energy mandates in the nation. Massachusetts, Connecticut, Rhode Island, New York, and Maine have each enacted legislation requiring dramatic reductions in greenhouse gas emissions and large-scale procurement of renewable energy, particularly offshore wind. These mandates create a fundamental tension with the wholesale market structures administered by the ISOs, which were designed around principles of fuel-neutral, least-cost competition. The question of how to reconcile state policy preferences with competitive market outcomes — a question that manifests most acutely in debates over capacity market rules, minimum offer price rules, and buyer-side mitigation — has been among the most contentious in American energy regulation over the past decade.

ISO-NE: The Forward Capacity Market and the Struggle for Resource Adequacy

ISO New England operates the bulk power system serving Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont — approximately 14.5 million people in a region that, despite its relatively modest geographic footprint, presents formidable operational challenges. The region has limited indigenous fuel resources, no significant domestic natural gas production, constrained pipeline capacity from production regions in Pennsylvania and the Gulf Coast, and a load profile dominated by the Boston metropolitan area and the Connecticut coastline.

The Forward Capacity Market and Pay-for-Performance

The centerpiece of ISO-NE's resource adequacy framework is the Forward Capacity Market (FCM), which procures capacity commitments three years in advance of the delivery period. The FCM was designed to provide a price signal sufficient to attract and retain the generation, demand response, and energy efficiency resources necessary to meet the region's reliability requirements. In theory, the forward structure gives developers sufficient lead time to finance and construct new resources; in practice, the FCM has been the site of persistent controversy over whether its prices are too high, too low, or distorted by out-of-market state subsidies.

A pivotal reform to the FCM was the introduction of the Pay-for-Performance (PFP) mechanism, which fundamentally altered the capacity market's incentive structure. Under PFP, capacity resources are not merely paid for being available in a general sense; they face significant financial penalties if they fail to perform during scarcity conditions and earn bonus payments if they over-perform. The design reflects a hard-learned lesson from the polar vortex events of 2013-2014, when a disturbing number of capacity resources — particularly gas-fired generators that lacked firm fuel supply — failed to produce when called upon during extreme cold. PFP was intended to shift the risk of non-performance from consumers to generators, creating a powerful financial incentive for resource owners to invest in fuel assurance, dual-fuel capability, on-site fuel storage, or other measures to ensure actual delivery during the hours that

matter most.

The PFP design has had measurable effects. It has encouraged investment in dual-fuel capability (allowing generators to switch from gas to oil during gas curtailments), incentivized battery storage development, and sharpened the distinction between resources that can reliably deliver during stressed conditions and those that cannot. Critics, however, have argued that PFP's penalty structure is punitive, that it disadvantages intermittent renewable resources that cannot guarantee output during any specific scarcity event, and that it has contributed to premature retirement of older units whose owners are unwilling to bear the performance risk. The tension between PFP's reliability-focused design and the region's simultaneous pursuit of large-scale renewable integration remains unresolved.

The Mystic Generating Station and Cost-of-Service Agreements

Perhaps no single episode better illustrates the contradictions of northeastern grid management than the saga of the Mystic Generating Station. Mystic, a large gas-fired combined-cycle plant located in Everett, Massachusetts, is directly connected to the region's only large-scale LNG import terminal, the Everett facility operated by Constellation (formerly Exelon). When the plant's owner announced its intention to retire the facility — arguing that FCM revenues were insufficient to justify continued operation — ISO-NE determined that Mystic was essential for regional fuel security, particularly during winter cold snaps when pipeline gas becomes scarce and the Everett LNG terminal serves as a critical backup fuel source.

The result was a controversial cost-of-service agreement, approved by the Federal Energy Regulatory Commission, under which Mystic was retained in operation outside the normal competitive market framework. The plant's costs — including fuel, maintenance, and a return on investment — were socialized across New England ratepayers. The arrangement drew sharp criticism from multiple directions. Market purists objected that it undermined competitive price signals and set a dangerous precedent for out-of-market reliability contracts. Consumer advocates questioned whether the costs were justified. Environmental groups pointed out that it effectively subsidized continued fossil fuel operation in a region with aggressive decarbonization mandates. State regulators chafed at bearing costs for a federal reliability determination over which they had no direct authority.

The Mystic episode laid bare a structural gap in eastern market design: the capacity market was designed to compensate generators for being available to produce electricity, but it did not adequately value the fuel-supply-chain infrastructure — specifically, LNG import and storage capability — upon which winter reliability increasingly depended. It also highlighted the growing frequency with which ISO-NE has been compelled to intervene outside normal market mechanisms to address reliability concerns that the market itself does not price.

Winter Reliability and the LNG Dependency

New England's winter reliability challenge deserves particular emphasis, as it represents one of the most acute fuel-security vulnerabilities in the American power system. The region's natural gas pipeline network — principally the Algonquin, Tennessee, and Iroquois systems — was sized for a demand profile that did not anticipate the massive shift from oil and coal to gas-fired generation that occurred

between 2000 and 2020. During winter, pipeline capacity is largely consumed by local distribution companies serving residential and commercial heating demand, which holds firm transportation contracts. Gas-fired generators, many of which rely on interruptible or secondary transportation, find themselves curtailed precisely when electricity demand is highest.

The consequence is a recurring pattern of winter price spikes and operational stress. During the December 2017-January 2018 cold snap, wholesale electricity prices in New England briefly exceeded $200 per megawatt-hour, and the ISO was compelled to call upon oil-fired generation, emergency demand response, and imported power from neighboring regions. LNG imports through the Everett terminal played a critical role in maintaining system reliability, but at spot-market LNG prices that were orders of magnitude above normal pipeline gas costs. The region has, in effect, become a seasonal importer of a globally traded commodity — liquefied natural gas — to compensate for the inadequacy of its domestic pipeline infrastructure, a situation that introduces both cost volatility and supply-chain risk.

Proposed solutions have ranged from pipeline expansion (politically and regulatorily blocked by state-level opposition in Massachusetts and New York), to LNG import terminal expansion, to aggressive energy efficiency and electrification programs intended to reduce gas heating demand, to large-scale offshore wind and battery storage deployment intended to reduce the region's dependence on gas-fired generation altogether. None of these solutions is available quickly or cheaply, and the winter fuel-security problem remains a defining challenge for ISO-NE.

State Renewable Mandates and Offshore Wind

Against this backdrop of infrastructure constraint and fuel vulnerability, the New England states have enacted some of the nation's most aggressive renewable energy mandates. Massachusetts, Connecticut, and Rhode Island have collectively procured or authorized more than 5,000 megawatts of offshore wind capacity through a series of competitive solicitations, with projects planned in federal lease areas south of Martha's Vineyard, Block Island, and along the southern New England continental shelf. Maine has pursued floating offshore wind technology suited to its deeper coastal waters. These procurements, driven by state Renewable Portfolio Standards and specific legislative mandates, represent the largest coordinated offshore wind buildout in American history.

The integration of offshore wind at this scale presents both opportunities and challenges for ISO-NE. On the opportunity side, offshore wind's production profile — which tends to peak in winter and correlate inversely with solar generation — complements the region's load shape and partially addresses the winter reliability gap. On the challenge side, the transmission infrastructure necessary to deliver offshore wind to onshore load centers requires substantial investment in subsea cables and onshore grid reinforcements, much of it in densely populated coastal areas where permitting is difficult. The intermittent nature of wind generation also complicates capacity market participation under the PFP framework, raising questions about how offshore wind resources should be valued for reliability purposes and whether existing market rules inadvertently discourage their development.

NYISO: Serving the World's Most Demanding Load Pocket

The New York Independent System Operator faces many of the same challenges as ISO-NE — aging infrastructure, gas dependency, winter fuel security, ambitious state mandates — but with the added complexity of serving New York City, arguably the most demanding single load pocket in the world.

Zone J: The New York City Load Pocket

NYISO divides its footprint into eleven pricing zones, lettered A through K. Zone J — New York City — is a load pocket of extraordinary characteristics. It serves more than eight million people in a dense urban environment where virtually all electricity must be imported across a limited number of transmission interfaces from upstate New York, New Jersey (via PJM), New England (via ISO-NE), and in-city generation. The transmission constraints into Zone J have historically produced persistent price separation between the city and the rest of the state, with New York City consumers paying substantially higher wholesale electricity prices due to congestion costs.

The in-city generation fleet has historically included a mix of large steam turbines (many dating to the mid-twentieth century), gas-fired combustion turbines, combined-cycle plants, and — until recently — the Indian Point nuclear facility. The physical constraints on generation within the five boroughs are severe: land is prohibitively expensive, air quality regulations are stringent, fuel delivery is constrained, and community opposition to power plants is intense. The result is a load pocket that operates with thin reliability margins, where the loss of a single large generating unit or transmission facility can have outsized consequences for system security.

The Indian Point Retirement

The retirement of Indian Point Energy Center — a two-unit, approximately 2,000-megawatt nuclear plant located in Buchanan, New York, roughly 35 miles north of midtown Manhattan — stands as one of the most consequential generation retirements in the history of the American grid. Indian Point had for decades provided a substantial share of New York City's baseload electricity, and its closure (Unit 2 in April 2020, Unit 3 in April 2021) removed a large block of carbon-free, fuel-secure generation from a transmission-constrained load pocket.

The replacement of Indian Point's capacity and energy has relied on a combination of new gas-fired generation (including the 680-megawatt CPV Valley plant and the Cricket Valley Energy Center), increased imports from upstate and neighboring regions, transmission upgrades, demand response, energy efficiency, and a growing portfolio of distributed solar and battery storage. Critics have noted the irony that the retirement of a zero-emission nuclear plant — driven by a combination of safety concerns, political opposition, and unfavorable economics — has resulted in increased reliance on fossil-fueled generation in the near term, even as the state pursues aggressive long-term decarbonization targets. The episode underscores a recurring tension in clean energy transitions: the sequencing of retirements and replacements matters enormously, and the premature closure of existing zero-carbon resources can produce outcomes that are counterproductive from an emissions standpoint.

The Climate Leadership and Community Protection Act

New York's Climate Leadership and Community Protection Act (CLCPA), enacted in 2019, established

one of the most aggressive decarbonization frameworks of any American state. The law mandates 70 percent renewable electricity by 2030, a zero-emission electricity sector by 2040, and economy-wide net-zero greenhouse gas emissions by 2050. It also includes significant environmental justice provisions, requiring that at least 35 percent of the benefits of clean energy spending be directed to disadvantaged communities.

The CLCPA creates enormous pressure on NYISO's market design and planning processes. Meeting the 2040 zero-emission electricity target requires not merely adding large volumes of renewable generation — the state has procured or targeted 9,000 megawatts of offshore wind, 6,000 megawatts of distributed solar, and 3,000 megawatts of energy storage — but also managing the orderly retirement of the state's remaining fossil fleet without compromising reliability, particularly in the Zone J load pocket. The timeline is extraordinarily compressed by infrastructure standards: achieving a zero-emission grid within roughly fifteen years from the law's enactment requires a pace of renewable deployment, transmission construction, and fossil retirement that has no precedent in the state's history.

NYISO has responded with a series of market design initiatives, including proposals for a capacity market structure that would integrate the state's carbon-reduction goals with reliability requirements. These proposals have included consideration of a carbon pricing mechanism within the wholesale market — an approach that would, if implemented, represent a significant departure from the fuel-neutral market design that has characterized American ISOs and RTOs since their inception. The carbon pricing proposal has attracted both strong support (from those who argue that internalizing the cost of emissions within the wholesale market is the most efficient path to decarbonization) and strong opposition (from those who argue that it would raise costs for consumers, distort competitive market outcomes, and encroach on state regulatory authority).

Buyer-Side Mitigation and MOPR Controversies

Both ISO-NE and NYISO have been at the center of prolonged controversies over buyer-side mitigation rules and the Minimum Offer Price Rule (MOPR). These rules, in their various forms, were originally designed to prevent the exercise of buyer-side market power — that is, to prevent large load-serving entities from sponsoring below-cost new generation in order to suppress capacity market prices for their own benefit. In practice, however, the rules have increasingly collided with state clean energy policies.

The core tension is straightforward: when a state government provides financial support to a preferred resource — whether through renewable energy credits, zero-emission credits for nuclear plants, contracts for differences for offshore wind, or other mechanisms — that resource may be able to offer into the capacity market at a price below its true cost of new entry. Under strict MOPR rules, such resources would be required to offer at an administratively determined price floor, effectively preventing them from clearing the capacity market and earning capacity revenues. The result is that consumers pay twice: once for the state-mandated clean energy resource (through above-market contract costs) and again for the conventional capacity that clears the market in its place.

In NYISO, the buyer-side mitigation rules became a significant obstacle to the state's clean energy agenda, as they threatened to impose offer floors on state-subsidized renewable and nuclear resources. The debate produced years of litigation before the Federal Energy Regulatory Commission and the

federal courts, and was ultimately overtaken by broader FERC action to reform MOPR rules nationwide. In ISO-NE, analogous debates centered on the treatment of state-sponsored renewable resources in the FCM, with proposals ranging from strict MOPR application to the creation of separate "substitution auctions" that would allow state-sponsored resources to replace conventional capacity on a one-for-one basis.

The MOPR controversies in the Northeast crystallize a fundamental governance tension in the American electricity sector: the boundary between federal jurisdiction over wholesale markets and state authority over resource procurement and environmental policy. This tension, examined in greater detail in Parts V and VI of this volume, remains one of the most consequential unresolved questions in American energy law.

* * *

II. MISO and SPP: The Wind Belt Challenge

The Interior Grid: A Portrait of Abundance

If the northeastern grid is defined by constraint — constrained transmission, constrained fuel supply, constrained siting — the grids of the American interior are defined by abundance. The Great Plains and Upper Midwest contain wind resources of staggering scale. The National Renewable Energy Laboratory has estimated that the technical potential for onshore wind generation in the states covered by MISO and SPP exceeds the total electricity consumption of the United States many times over. Solar resources, while less exceptional than those of the Desert Southwest, are nonetheless substantial across the southern portions of both footprints. And unlike the Northeast, where every acre of potential generation or transmission siting is contested, the interior offers vast expanses of agricultural land where wind turbines and solar arrays can coexist with farming and ranching, and where transmission rights-of-way, while still challenging to acquire, face less intensive land-use competition.

The challenge of the interior grid is not scarcity but distance. The best wind resources are concentrated in western Iowa, Minnesota, the Dakotas, Nebraska, Kansas, Oklahoma, and the Texas Panhandle — areas with sparse population and low electricity demand. The major load centers — Chicago, Detroit, Indianapolis, Minneapolis-St. Paul, St. Louis, Houston, New Orleans — are hundreds of miles to the east and south, connected by a transmission network that was designed to serve a fundamentally different pattern of generation. Building the transmission infrastructure to bridge this gap — to move tens of thousands of megawatts of renewable energy from the Plains to the cities — is arguably the single most important infrastructure challenge facing the American electricity sector, and it is in MISO and SPP that this challenge is most acute.

MISO: The Continent's Crossroads

The Midcontinent Independent System Operator is, by geographic footprint, the largest RTO in the United States, stretching from Manitoba and the Dakotas through the Upper Midwest and down the Mississippi Valley to the Gulf Coast of Louisiana, Mississippi, and Texas. It serves approximately 45 million people across fifteen states and the Canadian province of Manitoba, and its generation portfolio includes a diverse mix of coal, natural gas, nuclear, wind, and a rapidly growing fleet of solar resources.

The Multi-Value Project Portfolio

MISO's most celebrated contribution to American transmission policy is the Multi-Value Project (MVP) portfolio, a set of seventeen long-distance transmission projects approved in 2011 with a combined estimated cost of approximately $6.7 billion. The MVP portfolio was designed to accomplish multiple objectives simultaneously: relieving congestion on heavily loaded transmission interfaces, enabling the delivery of wind energy from the western portions of the MISO footprint to load centers in the east, meeting state renewable portfolio standard requirements, and improving system reliability.

The genius of the MVP concept lay in its approach to cost allocation. Rather than attempting to assign the costs of each individual transmission project to the specific beneficiaries of that project — an approach that had historically produced endless disputes and paralyzed transmission development — MISO allocated MVP costs broadly across its entire footprint on a postage-stamp basis, proportional to each zone's share of total system load. This approach was predicated on the finding that the portfolio's benefits — measured in terms of reduced congestion costs, reduced energy prices, improved reliability, and facilitated renewable energy delivery — were broadly distributed across the system and exceeded the costs by a ratio of approximately 2.2 to 1.

The MVP portfolio has been widely regarded as a success, both in its physical outcomes (the projects have been largely completed and have delivered measurable congestion relief and wind integration benefits) and as a model for regional transmission planning. It demonstrated that large-scale, long-distance transmission development is possible within the RTO framework when accompanied by a cost allocation methodology that distributes costs in rough proportion to benefits and when supported by state regulatory consensus. The MVP experience has informed subsequent MISO transmission planning efforts, including the ambitious Long-Range Transmission Planning (LRTP) initiative launched in the early 2020s.

The Seasonal Capacity Auction

MISO's approach to resource adequacy differs significantly from the forward capacity markets of the Northeast. Rather than procuring capacity three years in advance, MISO conducts a seasonal capacity auction — the Planning Resource Auction — that clears on an annual basis, and has transitioned toward a seasonal construct that recognizes the different reliability challenges of summer and winter. This design reflects the characteristics of MISO's resource mix and load profile: unlike New England, where winter fuel security is the binding constraint, MISO has historically faced its tightest reliability conditions during summer peak demand, though the growing penetration of wind (which tends to

produce less during summer afternoons) and the retirement of dispatchable coal and gas resources have begun to create reliability concerns across multiple seasons.

The seasonal auction design has come under increasing scrutiny as MISO's resource adequacy position has tightened. In the 2022 capacity auction, portions of the MISO footprint — particularly the northern and central zones — experienced capacity shortfalls, with prices spiking from near zero to the cost of new entry. The result was a dramatic increase in capacity costs for utilities in those zones and a sharp reminder that the energy transition's combination of renewable additions and thermal retirements can produce capacity gaps if not carefully managed. MISO has responded with reforms to its seasonal capacity construct, enhanced accreditation methodologies that more accurately reflect the reliability contribution of variable resources, and intensified transmission planning to ensure that capacity-rich areas of the footprint can deliver energy to capacity-short areas during stressed conditions.

The MISO South Seam

One of the most significant structural challenges within the MISO footprint is the so-called MISO South seam — the relatively weak transmission interface between the northern portion of the MISO system (centered on the Upper Midwest) and the southern portion (centered on Louisiana, Mississippi, and Arkansas). MISO South was integrated into the RTO in 2013, when the Entergy system — a large, vertically integrated utility spanning much of the Gulf Coast — joined MISO after years of regulatory proceedings.

The integration of MISO South created a single market spanning from Manitoba to the Gulf of Mexico, but the physical transmission system connecting north and south remains limited. The constrained interface between MISO Midwest and MISO South limits the ability to share resources across the full footprint, creating what are in effect two sub-regions with different resource adequacy positions, different congestion patterns, and different exposure to weather-related risks. The MISO South seam has been identified as a critical bottleneck in multiple transmission planning studies, and its reinforcement is a central element of MISO's long-range transmission planning agenda. Strengthening this interface would not only improve resource sharing between the two sub-regions but also facilitate the delivery of Gulf Coast solar resources to the Midwest and Upper Midwest wind resources to the South.

SPP: Wind Integration Champion of the Great Plains

The Southwest Power Pool, headquartered in Little Rock, Arkansas, operates the bulk power system across a fourteen-state footprint centered on the Great Plains — Kansas, Oklahoma, Nebraska, the Dakotas, and portions of surrounding states. SPP's footprint encompasses some of the finest wind resources on the continent, and the organization has earned a reputation as the leading wind integration success story in the American electricity sector.

Record-Setting Wind Penetration

SPP has repeatedly set records for wind energy penetration that would have been considered impossible by grid operators a generation ago. The RTO has recorded hours in which wind generation has supplied more than 90 percent of total system load — an extraordinary achievement that reflects both the quality of the region's wind resources and the sophistication of SPP's operational practices. On an annual basis, wind energy has grown to supply more than 30 percent of the electricity consumed within the SPP footprint, a figure that continues to rise as new wind projects enter commercial operation.

Managing these penetration levels has required significant innovation in forecasting, dispatch, and market design. SPP has invested heavily in wind forecasting capabilities, using sophisticated meteorological models to predict wind output across its footprint on time horizons ranging from minutes to days. Accurate forecasting is essential for managing the ramp events — rapid increases or decreases in wind output — that characterize wind generation on the Great Plains, where weather systems can produce swings of several thousand megawatts within a few hours.

The Integrated Marketplace

SPP's wholesale market — the Integrated Marketplace, launched in 2014 — was designed with wind integration as a central consideration. The market includes a day-ahead market, a real-time balancing market, and a transmission congestion rights market, with features specifically tailored to manage the variability and uncertainty of high wind penetration. These features include flexible ramping products, uncertainty reserves, and dispatch algorithms that can accommodate rapid changes in wind output without compromising system reliability.

The Integrated Marketplace has produced significant economic benefits for the SPP region, reducing wholesale electricity costs through more efficient dispatch, facilitating the integration of low-marginal-cost wind generation, and reducing the need for expensive out-of-market reliability interventions. SPP's market monitoring reports have consistently documented net benefits in the range of $1.5 to $2.5 billion per year relative to the pre-market bilateral trading arrangements they replaced.

Curtailment and the Low-Load/High-Wind Challenge

The very success of wind integration in SPP has created a new challenge: curtailment during periods of low load and high wind output. During spring and fall nights, when electricity demand is at its lowest and wind generation is often at its strongest, SPP can face conditions in which available wind output exceeds total system demand. In these conditions, wind generators must be curtailed — ordered to reduce output — to maintain system balance.

Curtailment imposes economic costs on wind developers (who lose revenue for every megawatt-hour they cannot produce), complicates the financing of new wind projects (as investors must account for curtailment risk in their return calculations), and represents a waste of a zero-fuel-cost, zero-emission resource. SPP has pursued multiple strategies to reduce curtailment, including transmission expansion to export surplus wind to neighboring regions, market reforms to improve the dispatch efficiency of the generation fleet, and coordination with neighboring systems (particularly MISO and the western interconnection) to facilitate broader geographic sharing of wind resources.

The curtailment challenge underscores a critical insight about renewable energy integration: adding

generation capacity is necessary but not sufficient. Without adequate transmission to deliver renewable energy to load centers and adequate flexibility in the conventional generation fleet to accommodate renewable variability, high penetration levels can produce diminishing returns. The solution is not less wind but more transmission, more storage, and more sophisticated market and operational tools — a lesson that is applicable far beyond the SPP footprint.

SPP Expansion and the Markets+ Competition

SPP has pursued an ambitious agenda of geographic and functional expansion, seeking to extend its market footprint westward into the Mountain West and Pacific Northwest. The RTO West proposal — an effort to incorporate utilities in Colorado, Wyoming, and other western states into SPP's full RTO framework — represents a significant potential expansion that would extend organized market benefits to a region that has historically relied on bilateral trading and utility-specific balancing.

This westward expansion effort, however, has encountered competition from an unexpected quarter: the California Independent System Operator's Extended Day-Ahead Market (EDAM). CAISO's EDAM proposal offers western utilities the benefits of organized day-ahead market participation without requiring full RTO membership — an attractive proposition for utilities and state regulators wary of ceding operational control to a regional organization. SPP has responded with its own day-ahead market offering, Markets+, designed to provide a competitive alternative to EDAM for western utilities seeking market participation on less comprehensive terms than full RTO membership.

The competition between SPP's Markets+ and CAISO's EDAM represents a fascinating market-design rivalry that may ultimately determine the organizational structure of the western electricity sector. The outcome will depend on a complex interplay of economic analysis, state regulatory preferences, governance concerns (particularly western states' wariness of California's dominant role in CAISO), and the practical operational benefits each platform can deliver. This competition is examined in greater detail in Chapter 10.

The Transmission Imperative: Building the Bridge from the Plains to the Cities

The overarching challenge facing both MISO and SPP — and indeed, the American electricity sector as a whole — is the construction of long-distance, high-voltage transmission infrastructure to connect the renewable energy resources of the Great Plains to the load centers of the Midwest, South, and East. The scale of the transmission buildout required is staggering. Studies by the National Renewable Energy Laboratory, the Department of Energy, and the RTOs themselves have consistently identified the need for tens of thousands of miles of new high-voltage transmission, at costs measured in tens of billions of dollars, to enable the level of renewable energy deployment envisioned by state mandates and federal policy goals.

The Barriers to Interstate Transmission

The barriers to interstate transmission development are formidable, and they are as much political,

regulatory, and institutional as they are engineering or financial.

Permitting. Unlike the interstate highway system or the natural gas pipeline network, the electric transmission system has no comprehensive federal siting authority. Transmission developers must obtain permits and rights-of-way from every state, county, and municipality through which a proposed line passes. A single 500-mile transmission project might require approvals from two or three state utility commissions, dozens of county governments, and hundreds of individual landowners. The process can take a decade or more, and a single adverse ruling at any point in the chain can delay or kill a project. Federal legislation granting backstop siting authority for transmission lines of national significance — analogous to the authority the Federal Energy Regulatory Commission holds for natural gas pipelines — has been debated for years but has never been enacted in a form that provides meaningful acceleration.

Cost allocation. Determining who pays for a new interstate transmission line is perhaps the most contentious question in American electricity regulation. The benefits of a transmission line accrue to multiple parties — generators who gain access to markets, consumers who gain access to lower-cost generation, system operators who gain reliability and flexibility — but these benefits are distributed unevenly across states and utility service territories. States that host a transmission line bear the costs of land use, visual impact, and construction disruption but may receive relatively little of the energy the line delivers. States that receive the energy benefit from lower prices but may be reluctant to pay for infrastructure located elsewhere.

The principle articulated by FERC — that costs should be allocated roughly commensurate with benefits — is sound in theory but ferociously difficult to apply in practice. Benefit-cost analyses for transmission projects involve assumptions about future fuel prices, generation development patterns, load growth, and policy trajectories that are inherently uncertain and easily contested. The result is that cost allocation disputes have delayed or blocked numerous transmission projects that would likely produce net benefits for the system as a whole.

State regulatory resistance. Even when permitting and cost allocation hurdles can be overcome, state regulators and legislatures may resist transmission projects that are perceived as exporting local resources (and associated economic benefits) to other states, or that are perceived as imposing costs on local ratepayers to benefit consumers elsewhere. This resistance is not irrational — it reflects the reality that state regulators are accountable to their own states' consumers and are not well-positioned to weigh the interests of a broader multi-state system. But it creates a collective action problem in which the regional or national benefits of transmission expansion are systematically undervalued relative to localized costs and political resistance.

Emerging Solutions

Despite these barriers, there are reasons for cautious optimism about the trajectory of transmission development in the interior. MISO's Long-Range Transmission Planning process has identified a portfolio of high-voltage transmission projects — the LRTP Tranche 1 portfolio, approved in 2022 at an estimated cost of approximately $10 billion — designed to accommodate the renewable energy growth and thermal generation retirements projected over the coming decades. This portfolio, building on the MVP precedent, uses a broad regional cost allocation methodology and has received support from a

coalition of state regulators, utilities, and clean energy advocates.

SPP has similarly advanced transmission planning efforts to accommodate its growing renewable portfolio, including interregional studies with MISO to identify opportunities for coordinated transmission investment across the seam between the two RTOs. The Joint Targeted Interconnection Queue (JTIQ) study, conducted collaboratively by MISO and SPP, identified a set of transmission projects at the MISO-SPP seam that would unlock significant quantities of wind and solar generation currently stuck in interconnection queues in both organizations.

Federal policy has also begun to shift. The Infrastructure Investment and Jobs Act of 2021 and the Inflation Reduction Act of 2022 included provisions for transmission financing, planning, and siting that, while falling short of comprehensive federal siting authority, represent the most significant federal commitment to transmission development in decades. The Department of Energy's designation of National Interest Electric Transmission Corridors and the creation of the Transmission Facilitation Program provide new tools for overcoming the barriers that have historically stymied long-distance transmission projects.

* * *

Conclusion: Convergent Challenges, Divergent Solutions

The four ISOs and RTOs examined in this chapter occupy opposite poles of the American grid challenge, yet their trajectories are converging. The northeastern systems, defined by urban density and infrastructure constraint, are pursuing massive offshore wind buildouts and transmission upgrades that will transform their resource mix from fossil-dependent to renewable-dominant — but only if they can solve the siting, permitting, and market-design challenges that have historically limited new infrastructure development. The interior systems, defined by renewable abundance and geographic scale, are pursuing the long-distance transmission buildout necessary to deliver their wind and solar resources to distant load centers — but only if they can overcome the cost allocation, permitting, and state regulatory barriers that have historically constrained transmission development.

Both sets of challenges ultimately converge on the same fundamental questions: How should the costs of the energy transition be distributed? Who decides what gets built, and where? How can wholesale markets designed for a fossil-fueled system be reformed to accommodate — and appropriately value — the variable, distributed, and policy-driven resources that will dominate the grid of the future? And how can a regulatory framework divided between federal and state authority, designed for an era of vertically integrated utilities and local generation, be adapted to govern a system that is increasingly regional and interconnected in its physical reality?

The answers emerging from these four regions — ISO-NE's Pay-for-Performance capacity market, NYISO's exploration of carbon pricing, MISO's multi-value transmission planning, SPP's wind

integration innovations — represent some of the most important experiments in American energy governance. Their successes and failures will not remain regional curiosities; they will shape the design of markets, the trajectory of transmission development, and the pace of decarbonization across the entire American power system.

The Northeast and Midwest corridors, for all their differences, share one final characteristic: they are regions where the tension between the physical reality of the grid and the institutional structures that govern it is most acute. It is in these regions that the American electricity sector is being forced to confront the gap between twentieth-century institutions and twenty-first-century challenges — and it is here that the outlines of a new institutional framework, however imperfect and contested, are beginning to emerge.

* * *

Key Concepts for Review:

- Forward Capacity Market (FCM) and Pay-for-Performance (PFP)
- Cost-of-service reliability agreements and their market implications
- Winter fuel security and LNG dependency in New England
- Zone J (New York City) as a constrained load pocket
- Indian Point retirement and sequencing risk in decarbonization
- Climate Leadership and Community Protection Act (CLCPA)
- Buyer-side mitigation and Minimum Offer Price Rule (MOPR) controversies
- Multi-Value Project (MVP) transmission portfolio and regional cost allocation
- Seasonal capacity auction design versus forward capacity markets
- The MISO South seam and interregional transfer constraints
- Wind curtailment during low-load/high-wind conditions
- SPP Integrated Marketplace design for high wind penetration
- Markets+ versus EDAM competition for western market organization
- Interstate transmission development barriers: permitting, cost allocation, and state regulatory resistance

* * *

Part V

Transformation and Resilience

Chapter 10: Decarbonization and the "Inverter-Based" Grid

Introduction: More Than a Fuel Switch

The decarbonization of the American electric grid is often described in deceptively simple terms — swap coal and gas for wind and solar, add some batteries, and the job is done. This framing, while politically convenient, obscures what is arguably the most profound transformation in the history of electric power engineering. What is underway is not merely a substitution of fuel sources. It is a fundamental re-engineering of the physical operating principles upon which the grid has relied for more than a century.

Since the days of Edison and Westinghouse, the electric grid has been a system built around large rotating machines. Generators at coal plants, nuclear stations, and hydroelectric dams all share a common physical characteristic: they contain massive metallic rotors spinning in precise synchrony with one another, locked together across thousands of miles by the laws of electromagnetism. This synchronous rotation is not an incidental feature of the grid — it is the grid, in a very real physical sense. The frequency of alternating current, the stability of voltage, the ability of the system to absorb sudden shocks without collapsing — all of these properties emerge from the physics of heavy objects spinning together in lockstep.

The new resources displacing these machines operate on entirely different principles. A solar panel has no moving parts whatsoever. A lithium-ion battery is an electrochemical device. A modern wind turbine, while it does spin, is typically decoupled from the grid by power electronics that convert its variable output into grid-compatible alternating current. These technologies connect to the grid not through the electromagnetic coupling of a spinning rotor, but through solid-state devices called inverters — semiconductor switches that synthesize an alternating current waveform from a direct current source. The industry refers to them collectively as inverter-based resources, or IBRs.

This distinction — synchronous machine versus inverter — may sound like an engineering technicality, but its implications ripple through every dimension of grid planning and operations. The

retirement of synchronous generators removes physical properties that grid operators have always taken for granted: rotational inertia, short-circuit strength, reactive power from spinning mass, and the natural tendency of interconnected machines to resist changes in frequency. These properties were never explicitly paid for or procured because they were simply inherent in the technology. They came free with every coal plant and nuclear station. Now, as those plants retire, system operators must find ways to replace not just their energy output, but their physical contribution to grid stability — contributions that many planners never had to think about because they were always just there.

Simultaneously, the energy transition is reshaping not only the supply side of the grid but its demand side and institutional architecture as well. Millions of small-scale resources — rooftop solar panels, home batteries, electric vehicles, smart thermostats — are proliferating at the grid's edge, behind the customer's meter, in a domain historically invisible to wholesale market operators. Federal policy, most notably FERC Order 2222, is attempting to bring these distributed resources into the wholesale market framework, creating a two-way grid in which consumers are also producers and the boundary between the bulk power system and the distribution network becomes porous.

This chapter examines both of these transformations and the policy landscape that surrounds them. We begin with the physics — what changes when the grid shifts from spinning mass to digital power electronics — and then turn to the institutional and market reforms necessary to integrate millions of distributed resources. We conclude with the broader decarbonization policy landscape, including the Inflation Reduction Act, the challenge of achieving the final increments of a zero-carbon grid, and the critical role of transmission expansion.

* * *

I. The Physics of the Transition: From Spinning Mass to Power Electronics

Synchronous Generators and the Nature of Grid Frequency

To understand what the grid loses as conventional generators retire, one must first understand how those generators work and why their physical characteristics matter.

A synchronous generator is, at its core, an electromagnet mounted on a shaft that spins inside a set of stationary wire coils. As the magnetized rotor turns, it induces an alternating voltage in the surrounding coils — this is the fundamental mechanism by which mechanical energy is converted to electrical energy. In North America, the rotor is designed to spin at a speed that produces alternating current at precisely 60 cycles per second, or 60 hertz. A two-pole generator achieves this at 3,600 revolutions per minute; a four-pole machine at 1,800 RPM.

The critical feature of synchronous generators is that they are electromagnetically coupled to one another through the transmission network. When two synchronous machines are connected to the same grid, the alternating magnetic fields they produce interact through the wires connecting them, creating forces that tend to pull their rotors into alignment. If one machine tries to speed up, the electromagnetic coupling exerts a restoring torque that slows it down and speeds up the others. If one tries to slow down, the coupling pulls it back. The result is that all synchronous generators on an interconnected grid rotate in precise lockstep — they are, in effect, one enormous distributed machine. The frequency of the grid is simply the shared rotational speed of this collective machine, expressed in electrical cycles per second.

This electromagnetic coupling is what gives the grid its inherent stability. When a large generator suddenly trips offline — a turbine blade failure, a boiler tube rupture, a lightning strike on a transmission line — the electrical load that generator was serving does not vanish. Instead, it is instantaneously redistributed among all the remaining synchronized machines. Each remaining generator slows down slightly as it picks up a share of the orphaned load. The grid frequency dips — from 60.00 hertz to perhaps 59.95 or 59.90 hertz — but it does so gradually, over seconds, because the rotors of all the remaining generators are heavy. They have enormous rotational inertia: the stored kinetic energy of massive steel cylinders spinning at high speed.

System Inertia: The Grid's Shock Absorber

Rotational inertia is the electrical grid's shock absorber. It is the property that determines how quickly frequency changes in response to a sudden imbalance between generation and load. The relationship is governed by the swing equation, a foundational expression in power systems engineering that relates the rate of change of frequency to the difference between mechanical power input and electrical power output, divided by the total rotational inertia of the system.

In practical terms, a grid with high inertia — many large synchronous machines spinning — will experience slow, manageable frequency deviations following a disturbance. Operators have seconds, sometimes tens of seconds, to detect the problem and dispatch additional generation or activate reserves. A grid with low inertia will experience the same power imbalance as a much faster frequency swing. The rate of change of frequency, commonly abbreviated as RoCoF, increases as inertia decreases. If RoCoF is too high, protective relays on generators and loads may trip before corrective actions can take effect, potentially triggering a cascading failure.

The numerical significance is worth appreciating. The Eastern Interconnection of North America — the vast synchronized grid stretching from the Rockies to the Atlantic — has historically operated with an inertia constant on the order of several hundred gigawatt-seconds. The loss of a single large generator, even a 1,300-megawatt nuclear unit, produces a frequency deviation so small and so gradual that most customers would never notice. But this comfortable margin exists only because thousands of synchronous machines are spinning simultaneously, each contributing its rotational energy to the collective buffer.

Inverter-Based Resources: A Different Physics

Solar photovoltaic panels, battery energy storage systems, and modern Type 3 and Type 4 wind turbines connect to the grid through fundamentally different means. Rather than an electromagnetically coupled rotating mass, they use power electronic inverters — devices built from semiconductor switches (typically insulated-gate bipolar transistors, or IGBTs, or increasingly silicon carbide MOSFETs) that rapidly switch direct current on and off to synthesize an alternating current waveform.

A solar panel produces direct current. A battery stores and discharges direct current. A Type 4 wind turbine generates variable-frequency alternating current that is first rectified to direct current and then re-inverted to grid-frequency alternating current. In all cases, the inverter is the interface between the resource and the grid. And crucially, an inverter has no rotating mass. It has no rotational inertia. It does not electromagnetically couple to other generators on the system. It is, in electrical terms, a fast-acting, software-controlled current source that can change its output in milliseconds — far faster than any mechanical system — but it contributes nothing to the grid's inertial energy store.

The vast majority of inverters deployed on the grid today operate in what is called grid-following mode. A grid-following inverter works by measuring the voltage and frequency of the grid at its point of connection and then injecting current in synchrony with that measured signal. It follows the grid's frequency rather than helping to set it. It relies on the grid having a stable voltage waveform — provided by synchronous machines or other sources — as a reference signal to lock onto. A grid-following inverter is, in essence, a parasite on the frequency stability provided by synchronous generators. It is a perfectly adequate technology when synchronous machines are abundant and system inertia is high. It becomes problematic when those machines retire and inertia declines.

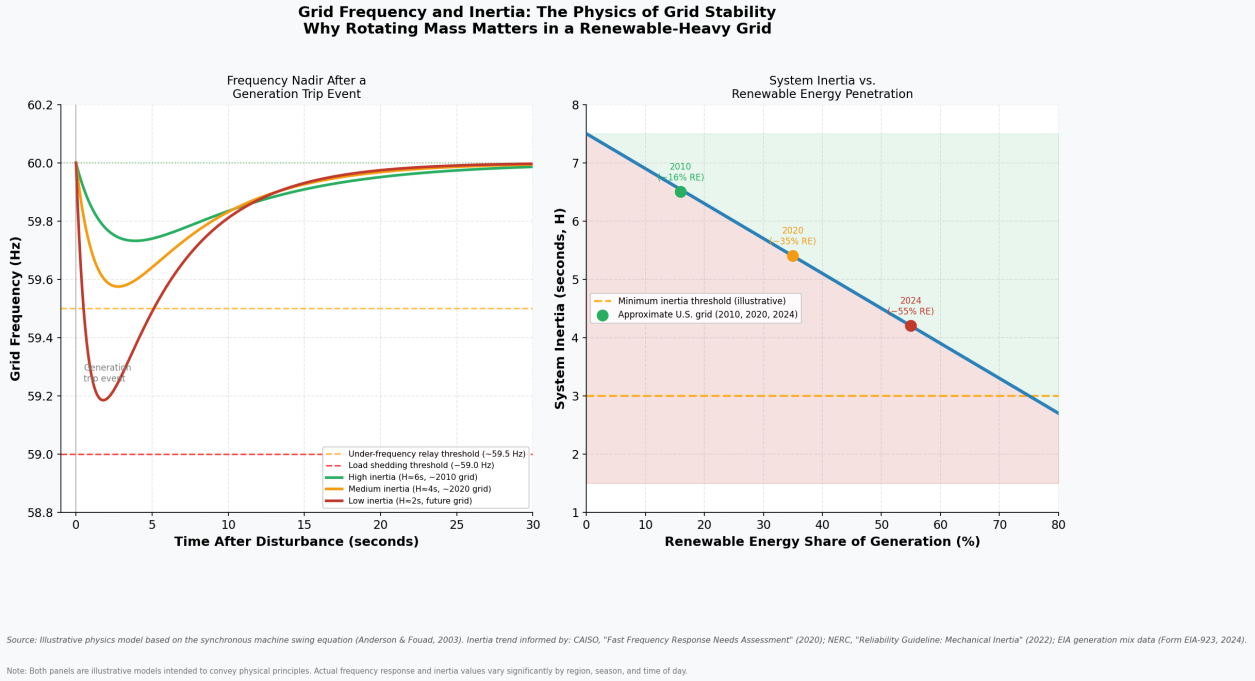

Figure 10.3: Grid Frequency and Inertia — The Physics of Grid Stability (Source: Anderson & Fouad, CAISO, NERC, EIA)

The Consequences of Declining Inertia

As thermal power plants retire across the United States and the share of inverter-based resources grows, system inertia is declining in measurable and operationally significant ways. The California Independent System Operator has documented a steady decrease in online synchronous inertia during spring afternoons, when solar generation is at its peak and gas-fired plants are backed down or decommitted. The Electric Reliability Council of Texas (ERCOT), which operates an islanded grid not synchronously connected to the rest of the country, has experienced periods where inverter-based resources supply more than 70 percent of total generation, pushing inertia to historically low levels.

The consequences are not hypothetical. On September 28, 2016, a severe storm in South Australia caused multiple transmission line faults in rapid succession. The resulting voltage disturbances triggered protection systems on several wind farms, which disconnected from the grid within milliseconds. The sudden loss of approximately 450 megawatts of wind generation — on a system with relatively low synchronous inertia because most conventional generators had already been retired or mothballed — caused frequency to plummet at a rate that overwhelmed the remaining generators' ability to respond. The entire state of South Australia blacked out. The subsequent investigation by the Australian Energy Market Operator identified low system inertia and the behavior of inverter-based resources during disturbances as key contributing factors.

Three years later, on August 9, 2019, Great Britain experienced a similar event. A lightning strike on

a transmission circuit caused the near-simultaneous tripping of a large offshore wind farm (Hornsea One) and a gas-fired generator (Little Barford). The combined loss of approximately 1,900 megawatts of generation on a system operating with relatively low inertia caused frequency to fall to 48.8 hertz — dangerously close to the 48.5-hertz threshold at which automatic load shedding begins. More than one million customers lost power. The investigation highlighted the rapid RoCoF and the unexpected tripping behavior of distributed solar inverters, which disconnected from the grid due to overly sensitive frequency and RoCoF protection settings, compounding the initial generation loss.

These events are harbingers of the challenges facing grid operators worldwide as the share of inverter-based resources continues to grow.

Grid-Forming Inverters: Teaching Power Electronics to Lead

The engineering community's primary response to the declining-inertia challenge is the development and deployment of grid-forming inverters. Unlike grid-following inverters, which measure and track an externally provided frequency signal, a grid-forming inverter establishes its own voltage and frequency reference. It behaves, from the grid's perspective, more like a synchronous machine — it acts as a voltage source behind an impedance rather than a controlled current source. When grid frequency drops, a grid-forming inverter can autonomously increase its power output in a manner analogous to the inertial response of a spinning rotor, drawing on whatever energy source backs it (a battery, a solar array with curtailment headroom, or a supercapacitor).

Grid-forming inverter technology is not science fiction. Several manufacturers now offer grid-forming battery energy storage systems, and pilot projects are operational in Australia, the United Kingdom, Hawaii, and parts of the continental United States. The technology is capable of providing synthetic inertia — a rapid power injection proportional to the rate of change of frequency that mimics the inertial response of a synchronous machine — as well as fast frequency response, voltage regulation, and black-start capability (the ability to energize a dead grid without an external power source).

The distinction between grid-forming and grid-following is sometimes described by analogy. A grid-following inverter is like a musician in an orchestra who watches the conductor and plays along. A grid-forming inverter is the conductor — or, more precisely, one of several conductors who must coordinate with one another. This coordination challenge is nontrivial. When multiple grid-forming inverters operate in proximity, their control algorithms must be designed to share load and stabilize voltage and frequency cooperatively, without the natural electromagnetic coupling that synchronizes rotating machines. This is an active area of research and standardization.

Other Solutions: Fast Frequency Response and Synchronous Condensers

Grid-forming inverters are not the only tool available. Battery energy storage systems, even those using grid-following inverters, can be programmed to provide fast frequency response — a rapid injection or absorption of power triggered by a detected frequency deviation. Because inverters can change their

output in milliseconds, battery-based frequency response is far faster than the mechanical governor response of a conventional generator, which typically takes several seconds to ramp. ERCOT has been a leader in deploying fast-responding battery resources for frequency regulation, and the results have been promising: battery storage can arrest frequency declines more quickly than traditional spinning reserves, partially compensating for reduced system inertia.

Another solution involves repurposing the physics of the old grid directly. A synchronous condenser is a synchronous machine — physically identical to a synchronous generator — that operates without a prime mover. It spins freely on the grid, providing rotational inertia, reactive power, and short-circuit strength without producing real energy. Some utilities have converted retired fossil fuel generators into synchronous condensers by disconnecting the steam turbine or gas turbine and allowing the generator rotor to spin as a motor/flywheel. Others have procured purpose-built synchronous condensers, often equipped with flywheels to increase their inertial contribution. This approach is particularly common in regions where the retirement of large thermal plants has created localized inertia deficits, such as parts of the Midwest and the Great Plains where wind penetration is high.

IEEE Standard 2800: Setting the Rules for IBR Performance

Recognizing that inverter-based resources are no longer marginal contributors to the grid but central participants, the Institute of Electrical and Electronics Engineers developed IEEE Standard 2800, first published in 2022. This standard, formally titled "IEEE Standard for Interconnection and Interoperability of Inverter-Based Resources Interconnecting with Associated Transmission Electric Power Systems," establishes uniform performance requirements for IBRs connecting to the bulk power system.

IEEE 2800 addresses many of the technical issues exposed by events like the South Australia blackout. It specifies ride-through requirements — the ability of an IBR to remain connected and continue operating during grid disturbances such as voltage sags, frequency excursions, and phase jumps. It establishes requirements for frequency and voltage support, reactive power capability, and power quality. It addresses the issue of IBR behavior during faults, requiring resources to inject reactive current to support grid voltage during transmission faults rather than tripping offline and compounding the disturbance.

The standard represents a philosophical shift in how the industry views inverter-based resources. Earlier interconnection standards treated IBRs primarily as energy sources that should disconnect from the grid when conditions become abnormal, protecting themselves at the expense of system stability. IEEE 2800 treats IBRs as essential grid participants that must contribute to system stability, much as synchronous generators have always been required to do. FERC and NERC have signaled their intent to incorporate IEEE 2800 requirements into mandatory reliability standards, making compliance a condition of interconnection for new IBR facilities.

* * *

II. FERC Order 2222 and the Distributed Grid

The Rise of Resources at the Grid Edge

While the bulk power system grapples with the physics of inverter dominance, an equally transformative change is occurring at the other end of the scale. Millions of small energy resources are proliferating behind the customer's electric meter — in homes, businesses, and communities — in a domain that the wholesale power market was never designed to see or manage.

Rooftop solar photovoltaic systems numbered more than four million installations in the United States by the mid-2020s. Residential battery storage systems, led by products like the Tesla Powerwall and the Enphase IQ Battery, are being installed at accelerating rates, driven by declining costs, state incentives, and customer desire for backup power during outages. Electric vehicles, each carrying a battery pack with 60 to 100 kilowatt-hours of storage capacity, represent a vast and growing fleet of mobile energy resources. Smart thermostats, connected water heaters, and controllable pool pumps offer flexible demand that can be shifted in time without affecting customer comfort.

These resources are collectively known as distributed energy resources, or DERs. Individually, each is small — a rooftop solar system might be 8 kilowatts, a home battery 13.5 kilowatt-hours. But in aggregate, they represent enormous capacity. The combined capacity of distributed solar in the United States exceeds that of many large conventional generators. The residential battery fleet, while still smaller, is growing rapidly. And the demand flexibility available from smart devices, if properly harnessed, could rival the output of dozens of peaking power plants.

The problem is that these resources have historically been invisible to the operators of the wholesale power markets. The RTOs and ISOs that run the day-ahead and real-time energy markets see the grid at the transmission level — they manage large generators, high-voltage transmission lines, and bulk load. What happens behind a customer's meter, on the distribution system, is the domain of the local distribution utility and the state public utility commission. This jurisdictional boundary, reinforced by decades of regulatory precedent, has effectively excluded distributed resources from participating in wholesale markets.

The Substance of Order 2222

On September 17, 2020, FERC issued Order 2222, a landmark rulemaking that requires all RTOs and ISOs to establish rules allowing aggregations of distributed energy resources to participate in their wholesale markets. The order was the culmination of years of incremental policy development and was explicitly framed as removing barriers to competition and ensuring that the wholesale markets benefit from the full range of available resources.

The core concept is aggregation. A single rooftop solar system or home battery is far too small to

participate meaningfully in a wholesale market that trades in megawatts. But a thousand home batteries, coordinated by a single aggregator and dispatched as a unified resource, can provide the same services as a small power plant — energy, capacity, frequency regulation, reserves. Order 2222 requires RTOs to create a participation model for these aggregations, allowing them to bid into energy, capacity, and ancillary service markets under rules comparable to those that apply to conventional generators.

The order defines distributed energy resources broadly. Any resource located on the distribution system or behind a customer meter qualifies, including solar, storage, demand response, electric vehicles (including vehicle-to-grid configurations), combined heat and power, fuel cells, microturbines, and other technologies. The order establishes a minimum size threshold of 100 kilowatts for an aggregation — far smaller than the typical minimum participation size in wholesale markets, which often ranges from 1 to 10 megawatts — though it allows RTOs to propose different thresholds with justification.

DER Aggregation in Practice: The Virtual Power Plant

The commercial and operational framework through which DER aggregation occurs is increasingly known as the virtual power plant, or VPP. A VPP is not a physical plant but a software platform and contractual structure that enables a single entity — the aggregator — to enroll, monitor, forecast, and dispatch thousands of distributed resources as if they were a single dispatchable generator.

The aggregator signs up individual DER owners (homeowners with batteries and solar, businesses with flexible loads, EV fleet operators) and installs or connects to communication and control hardware at each site. The aggregator's software platform continuously monitors the state of each resource — the battery's charge level, the solar array's output, the thermostat's setpoint — and aggregates this information into a portfolio-level view. When the RTO calls for energy or reserves, the aggregator dispatches signals to individual resources: discharge this battery, curtail that air conditioner, delay this EV's charging session. The RTO sees a single resource with a bid curve and a dispatch response; behind that single identity are hundreds or thousands of individual devices.

Several companies have demonstrated VPP capabilities at significant scale. Aggregators in ERCOT, PJM, and ISO-New England have enrolled tens of thousands of residential batteries and have dispatched them during peak demand events, grid emergencies, and for routine frequency regulation. In some cases, VPP dispatch during extreme heat waves or winter storms has demonstrably reduced wholesale market prices and avoided the need to call upon expensive peaking generation.

Technical Challenges of DER Market Participation

The technical challenges of integrating distributed resources into wholesale markets are substantial and not yet fully resolved.

Telemetry and metering present fundamental difficulties. Wholesale markets require real-time visibility into a resource's output and status, typically at four-second or ten-second intervals. Achieving this level of telemetry from thousands of residential devices — each connected through a home internet

connection of variable reliability — is a different engineering problem than monitoring a single large power plant with dedicated SCADA communications. The aggregator must handle data latency, communication dropouts, device malfunctions, and cyber-security vulnerabilities across a vast attack surface of consumer-grade hardware.

Settlement and verification are equally complex. When a VPP claims to have delivered 5 megawatts of demand reduction during a peak hour, how does the RTO verify that claim? Unlike a generator with a revenue-grade meter at its point of interconnection, a DER aggregation relies on inference — comparing actual consumption against a counterfactual baseline of what consumption would have been in the absence of the dispatch signal. Baseline estimation is inherently imprecise and subject to gaming. Metering individual DERs with revenue-grade accuracy is possible but expensive at scale.

Distribution system impacts create a layer of complexity that does not exist for bulk power resources. When an aggregator dispatches thousands of home batteries to inject power into the grid simultaneously, the resulting reverse power flows on distribution circuits can cause voltage violations, equipment overloads, and protection coordination problems. The distribution utility — which is responsible for the safe operation of its local network — has legitimate concerns about wholesale-dispatched DERs creating reliability problems on its system. Order 2222 acknowledges this by granting distribution utilities a limited role as a "relevant electric retail regulatory authority" (RERRA) that can review and potentially block DER participation if it would create reliability concerns on the distribution system. This gatekeeper role has been controversial, with aggregators arguing that utilities use reliability concerns as a pretext to exclude competitors, and utilities responding that aggregators do not understand or accept responsibility for distribution system constraints.

The Jurisdictional Tension

Order 2222 sits squarely atop one of the most contested jurisdictional boundaries in American energy regulation: the line between federal authority over wholesale markets and state authority over retail sales and distribution. FERC has clear jurisdiction over wholesale market rules, and it has the authority to require RTOs to open their markets to DER aggregations. But the DERs themselves — rooftop solar systems, home batteries, smart thermostats — are physically located on the distribution system, interconnected under state-jurisdictional rules, and often participating in state-jurisdictional programs (net metering, demand response, time-of-use rates) that may conflict with or be complicated by simultaneous participation in wholesale markets.

The order attempts to navigate this tension by allowing states to opt out — a relevant electric retail regulatory authority can prohibit DERs within its jurisdiction from participating in RTO markets through an aggregator if it determines that such participation conflicts with state law or policy. Several states have invoked or considered this opt-out, viewing federal DER market access as an encroachment on state regulatory authority. Others have embraced Order 2222 as complementary to their own distributed energy policies.

The compliance process has been slow and contentious. Each RTO was required to file tariff revisions implementing Order 2222, and these filings have been subject to extensive comment, protest, and revision. As of the mid-2020s, most RTOs have received FERC approval for at least initial compliance filings, but the details — minimum aggregation sizes, telemetry requirements, distribution utility review processes, dual participation rules — remain works in progress. Full, practical implementation of the DER aggregation participation model, with significant volumes of DERs actually bidding into and clearing in wholesale markets, remains more aspiration than reality in most regions.

The Vision of the Transactive Grid

Order 2222 and the VPP model represent early steps toward a more radical reimagining of the grid's architecture — sometimes called the transactive grid or the grid edge. In this vision, the grid evolves from a one-way system (large generators pushing power to passive consumers) into a networked platform in which millions of distributed resources actively participate in grid operations through automation and price signals.

Under a fully transactive model, a homeowner's battery would autonomously decide when to charge and discharge based on real-time wholesale prices, local distribution constraints, the homeowner's preferences, and weather forecasts. An electric vehicle would negotiate with the grid about when and how fast to charge, accepting a lower electricity price in exchange for flexibility about charging timing. A commercial building's energy management system would continuously optimize its HVAC, lighting, and on-site generation to minimize costs while providing grid services.

The enabling technologies for this vision are largely available: smart inverters, internet-connected devices, cloud computing platforms, machine learning algorithms for forecasting and optimization. The barriers are primarily institutional — market rules, regulatory frameworks, utility business models, consumer protection standards, and cybersecurity requirements that were all designed for a simpler, more centralized system. The transition from the current grid to a transactive grid is not primarily a technology challenge; it is a governance challenge, requiring the coordinated evolution of federal, state, and local regulatory frameworks.

* * *

III. The Decarbonization Policy Landscape

The Inflation Reduction Act and Clean Energy Economics

The Inflation Reduction Act of 2022 represents the most significant federal intervention in clean energy

economics in American history. Rather than mandating emissions reductions through regulation, the IRA operates primarily through the tax code, using production tax credits, investment tax credits, and bonus credit mechanisms to alter the economic calculus of energy investment.

The IRA extended and modified the Production Tax Credit (PTC), which provides a per-kilowatt-hour payment for electricity generated by qualifying resources over a ten-year period. For wind, solar, and other zero-emission generation, the base PTC value (adjusted for inflation) provides a significant and predictable revenue stream that substantially reduces the levelized cost of energy. The Investment Tax Credit (ITC), which provides an upfront credit equal to a percentage of a project's capital cost, was similarly extended and expanded, with particularly generous terms for energy storage — which for the first time received a standalone ITC, independent of pairing with solar generation.

Perhaps the IRA's most forward-looking provision is the introduction of technology-neutral clean energy tax credits, scheduled to take effect in 2025. These credits replace the technology-specific PTC and ITC structures with a unified framework that provides equivalent tax benefits to any zero-emission electricity generation technology. This technology-neutral approach is designed to avoid the historical pattern of Congress picking technological winners and losers, instead allowing the market to determine which zero-emission technologies are most cost-effective.

The IRA also introduced bonus credit mechanisms — additional credit value for projects that meet prevailing wage and apprenticeship requirements, for projects located in energy communities (areas with historical dependence on fossil fuel industries), for projects that use domestically manufactured components, and for projects located in low-income communities. These bonus provisions reflect a deliberate effort to align clean energy deployment with broader industrial policy and environmental justice goals.

The combined effect of the IRA's provisions has been a dramatic acceleration of clean energy investment. Utility-scale solar and battery storage project announcements surged in the years following the IRA's enactment, and manufacturing facilities for solar panels, battery cells, and other clean energy components have been announced at unprecedented scale. The law's long time horizons — many credits extend for a decade or more — provide the investment certainty that capital-intensive energy projects require.

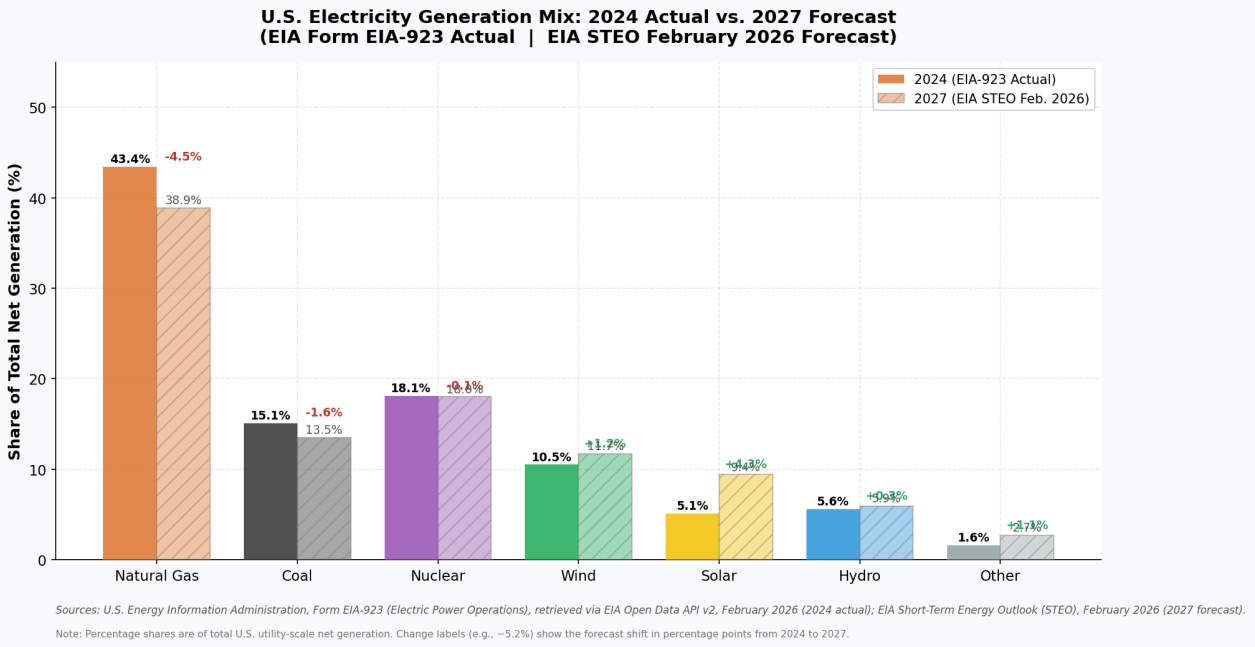

Figure 10.1: U.S. Electricity Generation Mix — 2024 Actual vs. 2027 Forecast (Source: EIA Form EIA-923, EIA STEO Feb. 2026)

State Renewable Portfolio Standards and Clean Energy Standards

Federal tax incentives operate alongside a patchwork of state-level policies that mandate or incentivize clean energy deployment. Renewable portfolio standards (RPS) — requirements that a specified percentage of electricity sold by utilities come from qualifying renewable sources — exist in more than thirty states and the District of Columbia. These standards vary widely in their ambition, timeline, and definition of qualifying resources. Some states set modest targets of 20 or 25 percent renewable energy by 2030. Others, including California, New York, and several New England states, have adopted 100 percent clean energy standards with target dates of 2040 or 2045.

The distinction between a renewable portfolio standard and a clean energy standard is substantive. An RPS typically qualifies only renewable resources — wind, solar, geothermal, some hydroelectric, some biomass. A clean energy standard (CES) qualifies any zero-emission or low-emission resource, potentially including nuclear power, carbon capture-equipped fossil generation, and other non-renewable but low-carbon technologies. Several states have transitioned from RPS to CES frameworks, recognizing that achieving very high levels of decarbonization may require a technology portfolio broader than renewables alone.

The interaction between state clean energy mandates and federal tax incentives creates a layered policy environment. In states with aggressive clean energy targets, the IRA's tax credits reduce the cost of compliance, effectively subsidizing the transition that state policy mandates. In states without clean energy mandates, the tax credits alone may be sufficient to drive substantial clean energy deployment on

purely economic grounds, as subsidized wind and solar are increasingly cheaper than operating existing fossil fuel plants.

The Last 10–20 Percent: The Hardest Part of Decarbonization

The early stages of grid decarbonization are, in relative terms, straightforward. Wind and solar are now the cheapest sources of new electricity generation in most of the United States, even without subsidies. Adding variable renewable energy up to penetration levels of 50 or even 60 percent of annual energy is technically achievable with existing technology, manageable amounts of battery storage, and incremental upgrades to grid operations. The challenge sharpens dramatically as decarbonization approaches 80, 90, or 100 percent.

The difficulty is rooted in the temporal variability of wind and solar resources. The sun does not shine at night, and both solar and wind output can be low for extended periods — days or even weeks — during certain weather patterns. As the grid approaches very high shares of variable renewable energy, the periods of deficit become the binding constraint. Four-hour lithium-ion batteries, the workhorse of current storage deployment, are well-suited to shifting solar energy from afternoon to evening but cannot bridge a week-long wind and solar drought in January.

Addressing this challenge requires what the industry calls firm clean energy — zero-emission resources that can generate electricity on demand, regardless of weather, for extended periods. Several candidate technologies are under development.

Long-duration energy storage encompasses a range of technologies designed to store energy for tens or hundreds of hours rather than the four to eight hours typical of lithium-ion systems. These include iron-air batteries, which use the reversible rusting of iron to store energy at low cost for very long durations; flow batteries, which store energy in liquid electrolytes in external tanks that can be scaled independently of power capacity; compressed air energy storage, which stores energy by compressing air into underground caverns; and gravitational storage systems, which lift heavy masses during charging and lower them to generate electricity during discharge. None of these technologies has achieved the scale or cost reduction of lithium-ion, but several are in pilot or early commercial deployment.

Advanced nuclear power, including small modular reactors (SMRs) and advanced reactor designs using molten salt, high-temperature gas, or liquid metal coolants, offers firm, dispatchable, zero-emission generation. SMRs are designed to be factory-manufactured and deployed in smaller increments than traditional large nuclear plants, potentially reducing construction risk and capital cost. Several SMR designs are in the licensing process with the Nuclear Regulatory Commission, and demonstration projects are underway. However, the nuclear industry carries the legacy of massive cost overruns and schedule delays from recent large reactor projects, and public acceptance remains a significant obstacle in many communities.

Green hydrogen — hydrogen produced by electrolyzing water using renewable electricity — represents another pathway to firm clean energy. Hydrogen can be stored in large quantities (in underground salt caverns, for example) and converted back to electricity through fuel cells or

combustion turbines when needed. The round-trip efficiency is relatively low (roughly 30 to 40 percent from electricity to hydrogen and back to electricity), making it an expensive form of energy storage. But hydrogen's advantage is that it can be stored in virtually unlimited quantities for very long durations at comparatively low marginal cost, making it potentially valuable for seasonal storage — absorbing surplus renewable energy in spring and summer and converting it back to electricity during winter peaks.

Carbon capture and sequestration (CCS) applied to natural gas generation offers another approach: retaining the dispatchability of gas turbines while capturing and permanently storing the carbon dioxide emissions. CCS technology has been demonstrated at scale in industrial applications, but its application to power generation has been limited by high costs and the energy penalty of the capture process. The IRA substantially increased the Section 45Q tax credit for carbon capture, improving the economics, but whether CCS can be deployed at the scale and cost necessary to play a significant role in grid decarbonization remains an open question.

The honest assessment of the current technology landscape is that no single technology has yet demonstrated the ability to provide firm, zero-emission electricity at the scale and cost necessary to decarbonize the last 10 to 20 percent of the grid. It is likely that a portfolio approach — combining long-duration storage, advanced nuclear, green hydrogen, CCS, and demand flexibility — will be necessary, with the optimal mix varying by region depending on resource availability, existing infrastructure, and local costs.

Transmission: The Critical Bottleneck

No discussion of grid decarbonization is complete without confronting the single largest physical bottleneck constraining the transition: the inadequacy of the transmission system.

The United States has approximately 160,000 miles of high-voltage transmission lines, most of which were built to connect large thermal power plants to nearby load centers. The geography of clean energy resources is fundamentally different from the geography of the fossil fuel fleet. The best wind resources are concentrated in the Great Plains and offshore. The best solar resources are in the Southwest and Southeast. The largest loads are in coastal cities and industrial centers. Connecting abundant clean energy resources to distant load centers requires massive investment in new long-distance, high-voltage transmission infrastructure.

The need is well-documented. Studies by the National Renewable Energy Laboratory, Princeton University, and other institutions consistently find that achieving a predominantly clean grid requires a doubling or tripling of transmission capacity, at a cost of hundreds of billions of dollars. The Department of Energy's National Transmission Needs Study has identified specific corridors where transmission congestion constrains clean energy delivery and increases electricity costs.

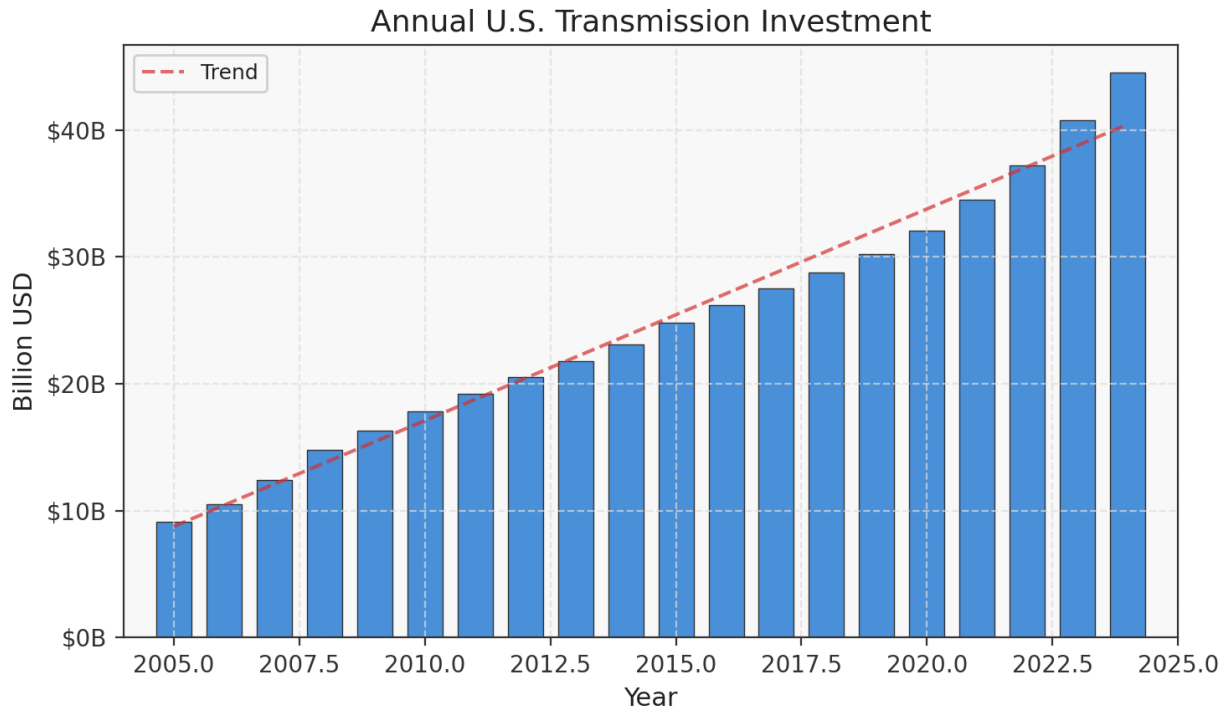

Figure 10.2: Annual U.S. Transmission Investment (Source: Edison Electric Institute, FERC Form 1)

The barriers to transmission construction are primarily institutional, not technological. Building a new long-distance transmission line requires navigating a thicket of federal, state, and local permitting requirements; acquiring rights-of-way across hundreds of miles of private land; resolving disputes over cost allocation (who pays for a line that crosses multiple states and benefits some more than others); and overcoming opposition from landowners, local governments, and environmental groups concerned about visual impact, land use, and ecological disruption. The average timeline for a major transmission project from conception to energization is ten to fifteen years, a pace fundamentally incompatible with climate timelines.

FERC has attempted to address the planning and cost allocation dimensions of this problem through Order 1920, issued in 2024. This order requires transmission providers to engage in long-range regional transmission planning on a forward-looking basis, considering anticipated future generation portfolios and load patterns rather than simply responding to interconnection requests after the fact. It also establishes principles for allocating the costs of regional transmission projects among beneficiaries, attempting to resolve the disputes that have historically stalled or killed major transmission proposals.

Order 1920 is a significant policy advance, but it addresses only the planning and cost allocation dimensions of the transmission bottleneck. It does not resolve the siting and permitting challenges that are often the most time-consuming and contentious aspects of transmission development. Federal permitting reform legislation has been debated in Congress, with proposals to streamline environmental review processes, establish federal backstop siting authority for nationally significant transmission lines,

and designate national interest electric transmission corridors. As of this writing, comprehensive permitting reform has not been enacted, and the transmission bottleneck remains the single greatest infrastructure constraint on the pace of decarbonization.

* * *

Conclusion: A Grid Rebuilt in Motion

The transformation described in this chapter is not a distant prospect — it is happening now, in real time, across every region of the American grid. Synchronous generators are retiring. Inverter-based resources are interconnecting at unprecedented rates. Distributed energy resources are proliferating behind the meter. Federal and state policies are pouring hundreds of billions of dollars into clean energy deployment. And the grid's operators, planners, and regulators are scrambling to adapt rules, markets, and operating procedures that were designed for a fundamentally different technological paradigm.

The challenge is often framed as a race — against climate change, against equipment obsolescence, against political cycles. But it is more accurately understood as an exercise in rebuilding an airplane in flight. The grid cannot be shut down for renovation. It must continue to deliver reliable power every second of every day while its physical foundations, market structures, and regulatory frameworks are simultaneously transformed.

The physics of the inverter-based grid are solvable. Grid-forming inverters, fast frequency response, synchronous condensers, and enhanced IBR performance standards provide a credible technical pathway to maintaining reliability as synchronous machines retire. The integration of distributed resources into wholesale markets, while institutionally complex, is achievable through the frameworks established by Order 2222 and the commercial model of virtual power plants. The economics of clean energy, supercharged by the Inflation Reduction Act, are increasingly favorable.

The binding constraints are institutional and infrastructural: the pace of transmission construction, the resolution of federal-state jurisdictional tensions, the adaptation of market designs conceived for a different era, and the development of firm clean energy technologies for the final and most difficult increment of decarbonization. These are challenges of governance, coordination, and collective investment — precisely the themes that have animated every chapter of this book. The grid has always been as much a political and institutional achievement as an engineering one. Its decarbonization will be no different.

* * *

Chapter 11: Cybersecurity and Physical Resilience

Introduction: The Infrastructure of Infrastructures

No single system built by human hands underpins modern civilization as thoroughly as the electric power grid. Water treatment plants require electricity to pump and purify. Telecommunications networks depend on it to route calls, transmit data, and power the cellular towers that knit together a wireless world. Hospitals, financial markets, military installations, traffic signals, fuel pipelines, natural gas distribution systems, food refrigeration supply chains — all of them presuppose, at every moment, the availability of electric power. The grid is not merely critical infrastructure. It is the infrastructure upon which all other infrastructure depends — the meta-infrastructure, the substrate, the foundation layer of technological society. When it fails, nothing else works for long.

This dependency has always existed, but two developments in the twenty-first century have elevated the protection of the grid from an engineering concern to a matter of national security. The first is the digitization of grid operations. The same supervisory control and data acquisition (SCADA) systems, energy management systems (EMS), and networked intelligent electronic devices (IEDs) that have made the grid more efficient and observable have also made it reachable — and therefore attackable — through cyberspace. A determined adversary sitting at a keyboard in Moscow, Beijing, or Tehran can now, in principle, reach into the control rooms of American utilities and open circuit breakers, disable protective relays, or corrupt the data upon which operators depend to make real-time decisions. The second development is climate change, which is reshaping the probability distribution of extreme weather events in ways that challenge the design assumptions embedded in every piece of infrastructure on the system. Hurricanes are intensifying more rapidly. Heat domes are persisting longer. Polar vortex disruptions are sending Arctic air into regions whose energy infrastructure was never designed for sustained, extreme cold. Wildfires are burning with an intensity and geographic reach that would have seemed implausible a generation ago.

This chapter examines both categories of threat — cyber and physical — and the frameworks that

have emerged to address them. These are fundamentally different kinds of risk, requiring different analytical tools and institutional responses. Cyber threats are adversarial and adaptive; the attacker learns and evolves in response to defenses. Physical threats from extreme weather are stochastic and intensifying; they do not adapt to defenses, but they are growing in frequency and severity in ways that render historical experience an unreliable guide to the future. What the two categories share is an uncomfortable truth: the consequences of failure are catastrophic, the investments required for protection are enormous, and the question of how to allocate costs and responsibilities remains deeply contested.

* * *

I. NERC CIP Standards and the Protection of Critical Infrastructure

The Institutional Framework: From Voluntary Reliability to Mandatory Security

For most of the history of the North American electric grid, reliability standards were voluntary. The North American Electric Reliability Council — note the word "Council," not "Corporation" — was formed in 1968 in the aftermath of the great Northeast blackout of 1965, and for nearly four decades it operated as a voluntary industry organization. Utilities participated because reliability was in their mutual self-interest, but no entity had the legal authority to compel compliance or impose penalties for violations. The system functioned tolerably well under this regime, in large part because the vertically integrated utility model aligned incentives: the same company that owned the generators, transmission lines, and distribution network had every reason to keep them operating reliably, because it bore the consequences of failure.

Two developments shattered this equilibrium. The restructuring of wholesale electricity markets in the 1990s and early 2000s disaggregated the system, creating new interfaces, new entities, and new coordination challenges that voluntary standards struggled to address. And the September 11, 2001 attacks — followed by growing awareness of cyber threats to industrial control systems — made the vulnerability of the grid a matter of urgent national security concern. The voluntary model was no longer adequate.

Congress responded with the Energy Policy Act of 2005, which authorized the Federal Energy Regulatory Commission (FERC) to certify an Electric Reliability Organization (ERO) with the power to develop and enforce mandatory reliability standards, including standards for the protection of critical infrastructure against cyber and physical threats. In 2006, FERC certified the North American Electric Reliability Corporation (NERC) — now a "Corporation" rather than a "Council," reflecting its new legal authority — as the ERO. This was a watershed moment in the governance of the grid. For the first time,

reliability standards were legally enforceable, backed by the power of the federal government, and violations could result in financial penalties of up to one million dollars per violation per day.

NERC does not operate alone. It delegates monitoring, compliance assessment, and enforcement to six Regional Entities — organizations like the ReliabilityFirst Corporation, the SERC Reliability Corporation, the Western Electricity Coordinating Council (WECC), and others — that serve as the front-line enforcers of NERC standards within their geographic territories. These Regional Entities conduct audits, review self-reports of violations, investigate potential noncompliance, and recommend penalties, subject to NERC and ultimately FERC oversight. The result is a layered compliance architecture: FERC sets policy direction and approves standards, NERC develops standards and oversees enforcement, and Regional Entities conduct the day-to-day work of compliance monitoring.

The CIP Standards: Architecture and Evolution

The Critical Infrastructure Protection (CIP) standards are the subset of NERC reliability standards specifically addressed to cybersecurity and the physical security of critical cyber assets. They have evolved substantially since their initial adoption, moving through multiple versions as threats have grown more sophisticated and as implementation experience has revealed gaps and ambiguities.

CIP-002: BES Cyber System Categorization. The entire CIP compliance framework rests on a foundational question: which assets are critical? CIP-002 establishes the methodology by which registered entities must identify and categorize their Bulk Electric System (BES) Cyber Systems — the cyber assets that, if compromised, could affect the reliable operation of the bulk power system. Assets are classified into three impact categories: high, medium, and low. High-impact assets include the control centers that operate transmission networks and balance areas. Medium-impact assets include large generating stations (typically those exceeding 1,500 MW at a single plant), transmission substations operating at 500 kV or above (and certain lower-voltage substations that meet specific criteria), and other facilities whose loss could have significant reliability consequences. Low-impact assets encompass the remainder of the BES Cyber Systems that do not meet the thresholds for high or medium categorization.

This categorization is consequential because it determines the stringency of the security controls that must be applied. High-impact facilities face the most rigorous requirements across every CIP standard. Medium-impact facilities face somewhat less demanding but still substantial requirements. Low-impact facilities were, for several years, subject to minimal requirements — a gap that NERC has progressively narrowed as the threat landscape has evolved and as policymakers have recognized that even smaller facilities can serve as entry points for sophisticated adversaries.

CIP-003: Security Management Controls. This standard requires responsible entities to develop and implement comprehensive cybersecurity policies. For high- and medium-impact systems, these policies must address the full range of security domains covered by the remaining CIP standards. For low-impact systems, CIP-003 requires, at a minimum, a cybersecurity awareness program, physical and electronic access controls, and incident response procedures. The standard ensures that cybersecurity is

treated as a matter of organizational governance, not merely a technical concern.

CIP-004: Personnel and Training. Human beings remain the most persistent vulnerability in any security architecture. CIP-004 requires entities to conduct personnel risk assessments, including criminal background checks, for individuals with authorized electronic or unescorted physical access to BES Cyber Systems. It mandates cybersecurity training programs and requires that access privileges be reviewed and revoked promptly when personnel change roles or depart the organization. The standard reflects a core principle of security engineering: access should be granted on the basis of need, limited to the minimum necessary for each role, and revoked the moment it is no longer required.

CIP-005: Electronic Security Perimeters. This standard requires entities to define and enforce Electronic Security Perimeters (ESPs) around their BES Cyber Systems — logical boundaries that control and monitor all electronic access. All traffic crossing an ESP must pass through identified Electronic Access Points (EAPs), and entities must implement controls to detect and prevent unauthorized access. CIP-005 has been updated to address the growing prevalence of remote access, interactive remote sessions, and vendor connectivity — attack vectors that have taken on heightened significance as utilities have adopted remote operations capabilities, a trend accelerated by the COVID-19 pandemic.

CIP-006: Physical Security of BES Cyber Systems. Cybersecurity and physical security are inseparable. A sophisticated firewall is worthless if an adversary can walk into an unlocked control room and plug a USB drive into a critical workstation. CIP-006 requires entities to define Physical Security Perimeters (PSPs) around the physical locations housing high- and medium-impact BES Cyber Systems, implement physical access control systems, maintain visitor logs, and monitor for unauthorized physical access.

CIP-007: System Security Management. This standard addresses the security configuration and management of the cyber systems themselves — logical access controls (authentication, authorization), security patch management, malware prevention, security event monitoring, and the management of system access accounts. The patch management requirement is particularly challenging in the operational technology (OT) environment, as we will discuss below, because patching often requires system downtime that is difficult to schedule on infrastructure that must operate continuously.

CIP-008: Incident Reporting and Response Planning. Entities must develop and maintain Cyber Security Incident Response Plans, test those plans through drills and exercises, and report qualifying cybersecurity incidents to the Electricity Information Sharing and Analysis Center (E-ISAC), which serves as the sector's primary hub for threat intelligence. Timely incident reporting serves both the affected entity and the broader sector, enabling collective defense by ensuring that information about new attack techniques and indicators of compromise is shared rapidly.

CIP-009: Recovery Plans for BES Cyber Systems. This standard requires entities to develop recovery plans specifying the procedures for restoring BES Cyber Systems following a cybersecurity incident. Plans must address backup and storage of information needed for recovery, testing of backup media, and the preservation of data that may be needed for forensic analysis. The standard reflects the recognition that no defense is impenetrable; the ability to recover quickly and restore operations is itself a critical security capability.

CIP-010: Configuration Change Management and Vulnerability Assessments. Entities must establish baselines for the configuration of their BES Cyber Systems and manage changes to those configurations through a documented process. They must also conduct periodic vulnerability assessments — at least once every fifteen months for high- and medium-impact systems — to identify and address security weaknesses before adversaries can exploit them.

CIP-011: Information Protection. BES Cyber System Information — documentation, network diagrams, configuration data, and other information that could be used to plan or execute an attack against the grid — must be identified, classified, and protected throughout its lifecycle, including during storage, transit, and disposal.

CIP-013: Supply Chain Risk Management. Adopted in 2020, CIP-013 represents one of the most significant expansions of the CIP framework. It requires entities to develop and implement plans for managing cybersecurity risks in their supply chains — the vendors, manufacturers, and service providers that supply hardware, software, and services for BES Cyber Systems. The standard was prompted by growing recognition that adversaries may find it easier to compromise a vendor's product or software update mechanism than to attack a utility directly. As the SolarWinds incident would dramatically illustrate, supply chain compromise represents a particularly insidious threat vector because it exploits trust relationships that are fundamental to the functioning of complex technological systems.

The Compliance and Enforcement Framework

NERC's compliance framework operates through a combination of scheduled audits, spot checks, self-reporting, and investigations triggered by events or complaints. Registered entities are required to maintain evidence of compliance — documentation, logs, records of assessments and training — and to produce that evidence upon request. Regional Entities conduct on-site and off-site audits on cycles determined by risk assessment, with higher-risk entities audited more frequently.

When violations are identified, the enforcement process provides for a range of dispositions. Minor violations posing minimal risk to the reliability of the bulk power system may be resolved through streamlined processes, including Find, Fix, Track, and Report (FFT) dispositions that emphasize remediation over punishment. More serious violations, particularly those involving intentional noncompliance or significant risk, can result in financial penalties up to the statutory maximum of one million dollars per violation per day. In practice, penalty amounts are determined by considering the severity and risk of the violation, the entity's compliance history, the entity's size and ability to pay, and the degree of cooperation in the investigation and remediation process. Aggregated penalties for serious violations have, in some cases, reached into the millions of dollars.

The enforcement framework serves a dual purpose. The threat of penalties creates incentives for compliance. But the standards themselves — the process of identifying critical assets, defining security perimeters, implementing access controls, conducting vulnerability assessments, and developing response and recovery plans — force a systematic, disciplined approach to cybersecurity that many entities would not undertake absent the regulatory mandate. The CIP standards do not, and cannot,

guarantee security. But they establish a baseline of security hygiene that raises the cost and difficulty of a successful attack.

Real-World Cyber Threats to the Grid

The CIP standards exist because the threats are real, present, and growing. Several incidents illustrate the nature and severity of the cyber threat to electric power systems.

The Ukraine Grid Attacks (2015 and 2016). On December 23, 2015, attackers — subsequently attributed by multiple intelligence agencies and cybersecurity firms to a Russian state-sponsored group known as Sandworm — successfully compromised the information technology and operational technology networks of three Ukrainian regional electricity distribution companies. The attackers used spear-phishing emails to gain initial access to the corporate IT networks, then moved laterally over a period of months to reach the SCADA systems controlling the distribution grid. On the day of the attack, they remotely operated circuit breakers to disconnect substations, used KillDisk malware to wipe systems and prevent operators from restoring control, flooded the utilities' call centers with denial-of-service attacks to prevent customers from reporting outages, and disabled uninterruptible power supply systems at the control centers. Approximately 225,000 customers lost power for periods ranging from one to six hours.

One year later, on December 17, 2016, the same group attacked a transmission substation in Kyiv using a more sophisticated and automated malware framework known as Industroyer or CrashOverride. This malware was specifically designed to communicate with industrial control systems using the native protocols of the electric power industry — IEC 61850, IEC 104, OPC DA — enabling it to directly operate circuit breakers without requiring human interaction. The 2016 attack was shorter in duration and more limited in scope, but it represented a qualitative escalation in capability: the development of purpose-built, reusable malware for attacking electric power systems.

The Ukraine attacks were a watershed for the global electric power industry. They demonstrated conclusively that cyber attacks on the grid were not theoretical but operational, not hypothetical but demonstrated. They also revealed the characteristic pattern of sophisticated grid attacks: patient, months-long preparation; exploitation of the connection between IT and OT networks; targeting of both the control systems and the support systems (communications, backup power, operator interfaces) needed for recovery.

The SolarWinds Supply Chain Compromise. In December 2020, the cybersecurity firm FireEye disclosed that it had discovered a supply chain compromise affecting SolarWinds, a widely used provider of network management software. Attackers — attributed to Russian intelligence — had inserted malicious code into software updates for SolarWinds' Orion platform, which were then distributed to approximately 18,000 organizations through the normal software update process. Among the affected organizations were federal agencies, defense contractors, and entities in the energy sector. The SolarWinds incident demonstrated the devastating potential of supply chain attacks: by compromising a single trusted vendor, the attackers gained access to thousands of downstream organizations

simultaneously. It provided stark validation for the concerns that had motivated the adoption of CIP-013.

State-Sponsored Threat Groups. The intelligence community and cybersecurity researchers have identified multiple state-sponsored threat groups that have targeted or are positioned to target the U.S. electric grid. Russian groups, including Sandworm and Berserk Bear (also known as Dragonfly or Energetic Bear), have conducted reconnaissance and prepositioning operations against U.S. energy infrastructure. Chinese state-sponsored groups, including Volt Typhoon, have been identified by U.S. intelligence agencies as having penetrated the networks of critical infrastructure entities — including energy companies — and maintaining persistent access that could be activated during a geopolitical crisis, particularly one involving Taiwan. The Department of Homeland Security, the Cybersecurity and Infrastructure Security Agency (CISA), and the intelligence community have assessed that China, Russia, Iran, and North Korea all possess some capability to conduct cyber operations against U.S. critical infrastructure, with China and Russia representing the most sophisticated and persistent threats.

The IT/OT Convergence Challenge

The fundamental cybersecurity challenge in electric power systems — and in all industrial control system environments — arises from the convergence of two historically separate technology domains. Information technology (IT) systems — corporate networks, email systems, business applications — were designed in an environment where cybersecurity, while never fully adequate, was at least a recognized concern from the outset. Operating systems are patched regularly. Antivirus software is deployed. Network traffic is monitored. Systems are refreshed on cycles measured in years.

Operational technology (OT) systems — the SCADA systems, remote terminal units, programmable logic controllers, protective relays, and other devices that directly monitor and control the physical grid — were designed in an era when cybersecurity was scarcely contemplated. These systems were engineered for reliability, determinism, and longevity. They use specialized industrial protocols that were designed for efficiency and interoperability, not for authentication or encryption. Many of these devices are embedded systems with limited processing power, incapable of running security software. They have operational lifespans measured in decades — far longer than the IT systems to which they are increasingly connected. Patching an OT system is fundamentally different from patching an office computer: it may require taking a substation or generator offline, it must be tested extensively to ensure that the patch does not introduce latency or functional changes that could affect grid operations, and the consequences of a failed patch can be physical — not merely an inconvenience, but a threat to equipment and human safety.

The convergence of IT and OT networks — driven by the operational benefits of remote monitoring, data analytics, and centralized control — has created pathways that sophisticated adversaries can exploit to move from the relatively accessible IT environment into the operational technology environment where they can cause physical consequences. The Ukraine attacks followed precisely this pathway. Defending against this threat requires not only the technical controls specified in the CIP standards but also a deep organizational commitment to network segmentation, access control, continuous monitoring,

and the cultivation of cybersecurity expertise that spans both the IT and OT domains.

The Electricity Subsector Coordinating Council

Recognizing that the defense of the grid against sophisticated nation-state cyber threats exceeds the capacity of any single utility — or even of the regulatory framework alone — the electric power industry and the federal government have established the Electricity Subsector Coordinating Council (ESCC) as the principal public-private partnership for grid security. The ESCC brings together the chief executives of major electric utilities and trade association leaders with senior officials from the Department of Energy, the Department of Homeland Security, FERC, and the intelligence community. It provides a forum for sharing classified threat intelligence, coordinating response plans, conducting joint exercises, and aligning public and private sector efforts to protect the grid.

The ESCC reflects a pragmatic recognition that grid cybersecurity is neither a purely private sector responsibility nor a purely governmental one. Utilities own and operate the assets, possess the operational expertise, and must implement the defenses. The government possesses intelligence about the adversaries, offensive cyber capabilities that can impose costs on attackers, and the legal and diplomatic tools needed to address state-sponsored threats. Effective grid defense requires the sustained integration of these complementary capabilities.

* * *

II. Hardening the Grid Against Climate-Driven "Black Swan" Events

Resilience and Reliability: A Critical Distinction

The electric power industry has long organized its planning and operations around the concept of reliability — the ability to deliver power to customers in the quantities they demand, when they demand it, at acceptable frequency and voltage. Reliability, as traditionally understood, encompasses two dimensions: adequacy (having sufficient generation and transmission capacity to meet anticipated demand, including reserves for contingencies) and operating reliability (the ability to withstand sudden disturbances, such as the loss of a major generator or transmission line, without cascading failures).

The planning standards that operationalize reliability are deterministic and well established. The N-1 criterion requires that the system be able to withstand the loss of any single element — a generator, a transmission line, a transformer — without violating operating limits or losing load. The more stringent

N-1-1 criterion requires the system to withstand two sequential contingencies with time for operator intervention between them. These criteria have served the industry well for decades. They are tractable, testable, and provide clear, auditable standards against which system performance can be measured.

But the events of the twenty-first century have exposed a gap between reliability as traditionally conceived and a broader concept that has come to be called resilience. Reliability standards address the system's ability to perform under normal conditions and anticipated contingency events. Resilience addresses the system's ability to withstand and recover from high-impact, low-probability events that exceed design criteria — events that reliability standards were never intended to address.

The distinction is not merely semantic. A system can be fully compliant with all applicable reliability standards and still be catastrophically vulnerable to a major hurricane, a prolonged polar vortex, a multi-day heat dome, or a wildfire that destroys hundreds of miles of transmission and distribution infrastructure simultaneously. The N-1 criterion asks: can the system survive the loss of one element? A Category 5 hurricane does not ask the system to survive the loss of one element. It destroys dozens or hundreds of elements simultaneously, across a wide geographic area, while also damaging the transportation and communications infrastructure needed to support restoration.

The Catalogue of Catastrophe: Recent Extreme Weather Events

The case for investing in resilience is written in the record of recent disasters.

Hurricane Maria (2017). When Hurricane Maria struck Puerto Rico on September 20, 2017, it destroyed the island's electric grid almost entirely. The Puerto Rico Electric Power Authority (PREPA) reported that one hundred percent of its customers — approximately 1.5 million — lost power. The restoration process took nearly eleven months; some customers in remote areas did not have power restored for a year. The prolonged outage contributed to a public health catastrophe. Studies estimated excess mortality in the thousands. The disaster exposed decades of deferred maintenance, inadequate investment, and institutional dysfunction, but it also revealed a more fundamental vulnerability: a small, isolated island grid with no interconnections to neighboring systems, limited redundancy, and infrastructure that had been designed for conditions less severe than those Maria delivered.

Hurricane Harvey (2017) and Hurricane Ian (2022). Harvey demonstrated the grid's vulnerability to prolonged flooding, as substations and other ground-level infrastructure were inundated across the Texas Gulf Coast. Ian, which struck southwest Florida as a near-Category 5 hurricane, destroyed distribution infrastructure across a wide area and required one of the largest mutual aid mobilizations in the industry's history, with tens of thousands of restoration workers deployed from across the country.

The PG&E Wildfire Crisis. In California, Pacific Gas and Electric Company (PG&E) faced a series of catastrophic wildfires linked to its transmission and distribution equipment — the Camp Fire of 2018, which destroyed the town of Paradise and killed 85 people, being the most devastating. PG&E was found liable for starting the fire through the failure of a transmission hook on a nearly century-old tower. The resulting liabilities drove PG&E into bankruptcy. In response, PG&E and other California utilities adopted Public Safety Power Shutoff (PSPS) programs — the deliberate de-energization of power lines

during high fire-risk conditions. PSPS events have affected millions of customers, presenting a bitter paradox: the utility proactively shuts off the power to prevent its own infrastructure from starting fires, imposing costs and hardships on customers to mitigate a risk that the infrastructure itself creates.

Winter Storm Uri (2021). In February 2021, a polar vortex disruption sent Arctic air deep into the southern United States, producing temperatures far below the historical design basis for energy infrastructure across Texas and neighboring states. The cascading failure that followed was a textbook illustration of the interconnection between the electric grid and the natural gas system. Natural gas wellheads, gathering lines, and processing plants froze and lost power. Without gas supply, gas-fired generators — which by 2021 constituted the largest share of Texas's generating fleet — tripped offline. The loss of generation forced the Electric Reliability Council of Texas (ERCOT) to order rotating outages to prevent a complete system collapse. But the outages were not "rotating" in any meaningful sense; millions of Texans lost power for days in subfreezing temperatures. Hundreds died. The financial consequences were staggering, with wholesale electricity prices spiking to the system's $9,000 per megawatt-hour price cap and natural gas prices reaching extraordinary levels. Uri demonstrated that reliability planning predicated on historical weather patterns was inadequate to address the changing risk landscape, and that the interdependence of the electric and natural gas systems created vulnerabilities that neither industry had adequately addressed.

Extreme Heat Events. Prolonged heat domes across the western United States have repeatedly stressed the grid, pushing electricity demand to record levels while simultaneously degrading the capacity of both transmission lines (which lose capacity as ambient temperatures rise) and generation sources (thermal plants lose efficiency, hydroelectric output declines during drought, and even solar panel efficiency decreases at very high temperatures). The combination of extreme demand and degraded supply creates conditions ripe for load shedding and rolling blackouts.

The Tools of Physical Resilience

Enhancing the physical resilience of the grid requires investments across multiple dimensions — hardening infrastructure to withstand more severe conditions, building redundancy and flexibility into the system, and improving the capacity for rapid restoration.

Undergrounding Distribution Lines. Burying power lines underground eliminates their exposure to wind, ice, falling trees, and wildfire ignition risk. Underground lines are dramatically less susceptible to weather-related outages than overhead lines. However, undergrounding is extraordinarily expensive — typically five to ten times the cost of overhead construction per mile — and underground lines, when they do fail, are more difficult and time-consuming to locate and repair. Undergrounding also does not protect against flooding, which can damage underground vaults and equipment. For these reasons, undergrounding is typically pursued selectively, in areas of highest risk or highest consequence, rather than as a system-wide strategy.

Vegetation Management. Trees and vegetation are the single largest cause of weather-related power outages on the distribution system and a significant cause on the transmission system. Utilities

spend billions of dollars annually on vegetation management programs — trimming trees, clearing rights-of-way, and removing hazard trees. Enhanced vegetation management, which involves removing not only trees within the right-of-way but also tall trees outside the right-of-way that could strike power lines if they fall, has been adopted by several utilities, particularly in high fire-risk areas. These programs are effective but expensive and often contentious, as they involve removing trees on or near private property.

Transmission Structure Hardening. Replacing wood poles with steel or concrete structures, designing structures to higher wind and ice loading standards, and upgrading hardware and conductors can significantly improve the survivability of transmission lines in extreme weather. Florida utilities, following the devastating hurricane seasons of 2004 and 2005, invested heavily in transmission hardening, and the investment has yielded measurable improvements in performance during subsequent storms. The trade-off, as always, is cost: hardened structures are more expensive, and the incremental cost must be weighed against the probability and consequences of the events they are designed to withstand.

Flood Protection for Substations. Substations are concentrations of high-value, long-lead-time equipment — particularly large power transformers, which can take twelve to eighteen months to manufacture and deliver. Protecting substations against flooding through elevated construction, flood walls, submersible equipment, and the relocation of critical facilities out of flood plains is an important resilience investment, particularly as flood maps are being redrawn to reflect changing precipitation patterns and sea level rise.

Microgrids and Islanding Capability. Microgrids — localized energy systems capable of operating independently from the bulk power system — can maintain power to critical facilities (hospitals, emergency operations centers, water treatment plants, military installations) when the broader grid fails. The concept of "islanding" — intentionally disconnecting a portion of the grid and operating it autonomously using local generation — extends the microgrid concept to community scale. Advances in distributed energy resources, battery storage, and microgrid control systems have made these configurations increasingly practical, though regulatory and interconnection challenges remain.

Mobile Emergency Transformers. Because large power transformers are both expensive and slow to manufacture, the industry has developed mobile transformers — trailer-mounted units that can be deployed rapidly to replace damaged substation transformers on a temporary basis. The Department of Energy has supported the development of strategic reserves of emergency transformers, and utilities have invested in transformer sharing programs. These efforts mitigate one of the most critical single points of failure in the grid, although the diversity of transformer specifications (voltage ratings, impedance characteristics, cooling requirements) limits the interchangeability of units.

Mutual Aid Agreements. The electric utility industry operates the most extensive mutual aid network of any infrastructure sector. When a major storm strikes, utilities from across the country deploy restoration crews, equipment, and materials to the affected area. These deployments are coordinated through the Edison Electric Institute's mutual aid framework and through regional mutual assistance groups. The mutual aid system is remarkably effective — it routinely mobilizes tens of thousands of workers across state lines within days — but it has limitations. If multiple regions are affected

simultaneously, mutual aid resources are stretched thin. And mutual aid addresses restoration, not prevention; it gets the lights back on after the damage is done.

The Challenge of Non-Stationarity

Perhaps the most fundamental challenge in resilience planning is the problem of non-stationarity — the recognition that historical weather data no longer provides a reliable guide to future conditions. For most of the history of infrastructure engineering, designers could reasonably assume that the climate was statistically stationary: the probability distribution of extreme weather events in the future would resemble that of the past. A one-hundred-year flood was a flood whose severity was expected to be equaled or exceeded, on average, once per century, based on the historical record.

Climate change has invalidated this assumption. Precipitation patterns are shifting. Temperature extremes are becoming more extreme. The frequency of Category 4 and 5 hurricanes is increasing. The wildfire season is lengthening. Design criteria based on historical weather data may systematically understate the risks that infrastructure will actually face over its operational lifetime. This creates a planning dilemma of the first order: how does one design infrastructure for a future climate that is uncertain, when the infrastructure itself has a useful life measured in decades?

The emerging answer involves a combination of approaches: using climate models and scenario analysis rather than historical data alone to inform design criteria, building in greater margins of safety, designing for adaptability (so that infrastructure can be upgraded or hardened as conditions change), and accepting that some degree of residual risk is inevitable and must be managed through operational measures, insurance, and recovery capabilities rather than eliminated through design alone.

The Regulatory Challenge: Who Pays for Resilience?

Resilience investments present a vexing regulatory challenge. Traditional utility ratemaking is built on the concept of "used and useful" — the utility invests in infrastructure that is needed to serve customers, and the regulator allows the utility to recover its prudent costs through rates. The cost-benefit analysis for reliability investments is relatively straightforward: the investment prevents outages that would otherwise occur with quantifiable frequency and duration, and the benefits can be measured in terms of avoided outage costs.

Resilience investments are different. They protect against low-probability, high-consequence events. The expected value calculation — the probability of the event multiplied by its cost — may produce numbers that are modest relative to the investment required, because the probability of any single catastrophic event in any given year is low. But the consequences of the event, if it occurs, are devastating and may include loss of life. Traditional cost-benefit analysis struggles with this asymmetry. It also struggles with the problem of deep uncertainty — when the probability of the event itself is uncertain and contested, as it is with climate-driven extreme weather, the expected value calculation rests on a foundation of sand.

Several approaches to this regulatory challenge have emerged. Some states have established dedicated resilience programs with separate cost-recovery mechanisms. Some regulators have adopted performance-based incentives that reward utilities for resilience outcomes (such as faster restoration times) rather than specifying particular investments. FERC has initiated proceedings to consider whether resilience should be explicitly valued in wholesale markets and transmission planning. But no consensus framework has emerged, and the tension between the desire for resilient infrastructure and the imperative to keep rates affordable remains unresolved.

* * *

III. Electromagnetic Pulse and Geomagnetic Disturbance Threats

Geomagnetic Disturbances

The sun periodically produces coronal mass ejections (CMEs) — massive eruptions of magnetized plasma that, if directed toward Earth, can interact with the planet's magnetic field and induce geomagnetically induced currents (GICs) in long conductors on the Earth's surface, including electric transmission lines, pipelines, and communications cables. GICs flow through the grounded neutrals of high-voltage transformers, driving the transformer cores into saturation. A saturated transformer draws enormous reactive power from the system, generates damaging levels of heat in structural components not designed for such heating, and produces harmonic distortion that can trigger protective relay operations across the system.

The most severe geomagnetic disturbance in the modern era occurred on March 13, 1989, when a CME-driven geomagnetic storm caused the collapse of the Hydro-Quebec system within ninety seconds, blacking out the entire province of Quebec for approximately nine hours. The event damaged transformers and other equipment on systems as far south as New Jersey. In 2003, a severe geomagnetic storm damaged transformers in South Africa and caused operational disruptions on systems worldwide.

The historical record suggests that significantly more severe events are possible. The Carrington Event of 1859 — the most intense geomagnetic storm ever recorded — produced GICs so powerful that telegraph operators reported receiving shocks and observed sparks arcing from their equipment. A Carrington-class event today, striking a grid that is vastly more extensive and more dependent on sensitive electronic equipment than the telegraph network of 1859, could potentially damage hundreds of high-voltage transformers simultaneously, across a continent-wide footprint. Because large power transformers require twelve to twenty-four months to manufacture, simultaneous damage to a large number of transformers could result in a prolonged, widespread outage measured not in hours or days but in months or years.

NERC has addressed this threat through TPL-007, a standard requiring transmission owners and operators to conduct vulnerability assessments of their systems to geomagnetic disturbances and to develop corrective action plans for facilities that do not meet acceptable performance levels. The standard defines benchmark GMD events against which the system must be assessed and requires entities to install GIC monitoring equipment and to develop operating procedures for GMD events. Mitigation measures may include the installation of GIC blocking devices (neutral blocking capacitors or resistors that prevent GIC from flowing through transformer neutrals), the procurement of spare transformers for the most critical locations, and operational procedures for reducing system loading during GMD events to provide additional thermal margin.

Electromagnetic Pulse

An electromagnetic pulse (EMP) produced by the detonation of a nuclear weapon at high altitude represents perhaps the most severe single threat to the electric grid. A high-altitude nuclear detonation produces three distinct pulse components. The E1 pulse, occurring within nanoseconds, is a brief, intense electromagnetic field that can damage or destroy unhardened semiconductor-based electronic equipment — including the digital relays, control systems, and communications equipment on which modern grid operations depend. The E2 pulse, occurring over microseconds to milliseconds, resembles the electromagnetic effects of lightning and is generally within the capability of existing lightning protection to address. The E3 pulse, occurring over seconds to minutes, resembles a severe geomagnetic disturbance and can produce damaging GICs in long transmission lines.

The EMP threat occupies a contested space in policy debates. The Congressional EMP Commission, established in 2001 and reconstituted in 2015, has warned that a single high-altitude nuclear detonation over the center of the continental United States could produce an EMP affecting the entire nation, potentially causing a prolonged, nationwide blackout with catastrophic consequences for society. The Commission recommended significant investments in EMP hardening of critical grid infrastructure. Critics have argued that the Commission's worst-case scenarios are exaggerated, that the cost of comprehensive EMP hardening would be prohibitive, and that the probability of a high-altitude nuclear attack — which would constitute an act of war and invite nuclear retaliation — is extremely low.

The debate over EMP hardening illustrates, in acute form, the broader challenge of resilience planning: how much should society invest to protect against events that are catastrophic but rare, whose probability is deeply uncertain, and whose consequences are disputed among experts? There is no objectively correct answer to this question. It is ultimately a judgment about acceptable risk, made through political processes that must weigh the claims of resilience investment against competing demands on limited resources.

* * *

IV. Toward an Integrated Framework for Grid Security and Resilience

The cyber and physical threats discussed in this chapter differ in their mechanisms but converge in their implications. Both can cause widespread, prolonged outages. Both can cascade through the interdependencies that connect the electric grid to other critical infrastructure systems. Both require investments that exceed what any individual utility can or will undertake without regulatory mandates or incentives. And both demand institutional architectures that cross the traditional boundaries between the public and private sectors, between federal and state regulatory authority, and between the engineering disciplines that have historically operated in separate silos.

Several principles should guide the development of an integrated security and resilience framework. First, the framework must be threat-informed and risk-based, prioritizing investments that address the most consequential vulnerabilities rather than pursuing uniform hardening that treats all assets and all threats as equivalent. Second, it must be adaptive, recognizing that both cyber adversaries and climate risks are evolving and that static defenses and fixed design criteria will be progressively less adequate over time. Third, it must address interdependencies — particularly the critical interdependence between the electric grid and the natural gas system, which Winter Storm Uri exposed with such devastating clarity. Fourth, it must grapple honestly with the question of cost allocation, developing regulatory frameworks that enable necessary investments while distributing their costs equitably among the ratepayers, shareholders, and taxpayers who benefit from a secure and resilient grid.

The electric grid is the infrastructure that makes all other infrastructure possible. Its protection against the full spectrum of threats it faces — from the patient, sophisticated cyber operations of nation-state adversaries to the raw, indiscriminate destructive power of a Category 5 hurricane or a Carrington-class geomagnetic storm — is not merely an engineering challenge or a regulatory obligation. It is a civilizational imperative.

* * *

Key Concepts: NERC CIP standards, BES Cyber Systems, Electronic Security Perimeters, IT/OT convergence, SCADA security, supply chain risk management, grid resilience vs. reliability, N-1 contingency criteria, undergrounding, vegetation management, transmission hardening, microgrids, mutual aid, geomagnetic disturbance (GMD), electromagnetic pulse (EMP), non-stationarity, climate adaptation, Electricity Subsector Coordinating Council (ESCC)

Part VI

The Digital Frontier

Chapter 12: Data Centers and the New Geography of Load

Introduction: The Electrification of Computation

In the closing decades of the nineteenth century and the opening decades of the twentieth, the American economy underwent a transformation so thoroughgoing that it is difficult, from our present vantage, to fully appreciate its magnitude. The electrification of industry — the replacement of steam engines, line shafts, and belt drives with electric motors — did not merely substitute one energy source for another. It reorganized the factory floor, enabled new manufacturing processes, altered the geography of industrial production, and ultimately rewired the relationship between energy and economic output. The electric grid, as described in the preceding chapters of this book, was built to serve this transformation. Its physical architecture, its regulatory institutions, its market structures, and its planning assumptions all reflect, in deep and often invisible ways, the patterns of demand that characterized the industrial and post-industrial American economy.

We are now in the early stages of a comparable transformation — one that, like the electrification of industry, threatens to overwhelm the assumptions embedded in the existing system. The rapid growth of data centers — vast facilities housing servers that store, process, and transmit the digital information upon which modern economic life depends — represents the emergence of a genuinely new category of electric load, one whose scale, characteristics, and geographic distribution differ fundamentally from anything the grid was designed to accommodate. After roughly a decade of essentially flat electricity demand growth in the United States — a period, spanning approximately 2010 to 2022, in which energy efficiency gains roughly offset the effects of population growth, economic expansion, and the proliferation of electronic devices — data centers are driving a sudden and dramatic resurgence in load growth that has caught utilities, grid planners, and regulators by surprise.

Data centers currently consume approximately four to five percent of total United States electricity generation, a figure that, while already substantial, understates the trajectory. Credible projections from utilities, consulting firms, and government agencies suggest that data center electricity consumption

could reach eight to twelve percent of the national total by 2030, driven in particular by the explosive growth of artificial intelligence workloads. In certain regions — most notably northern Virginia, central Ohio, and parts of Georgia and Texas — data center load growth is not merely significant but transformative, forcing utilities to revise integrated resource plans that were premised on decades of near-zero demand growth.

This chapter examines data centers not primarily as a technology story — the literature on server architecture, cloud computing, and artificial intelligence is vast and growing — but as a grid story. The central question is what this new category of demand means for the physical, economic, and institutional architecture of the American power system: the transmission lines and substations that must carry the power, the generators that must produce it, the markets that must price it, and the regulatory institutions that must govern the allocation of costs and risks among the various parties. The data center phenomenon touches every dimension of the grid that the preceding chapters have explored — from the physics of frequency regulation discussed in Chapter 1 to the market design challenges analyzed in Chapters 5 and 6 to the decarbonization imperatives examined in Chapter 10 — and introduces complications that existing frameworks are only beginning to address.

* * *

I. Anatomy of a Data Center: The Electrical Profile

Power Architecture and Reliability

To understand what data centers mean for the grid, one must first understand what a data center looks like from the perspective of the electric power system — not as a building full of servers, but as an electrical load with particular characteristics, requirements, and behaviors. The electrical architecture of a modern data center is, in its own way, as carefully engineered as the grid itself, and it reflects a set of priorities — above all, the absolute imperative of uninterrupted power — that have profound implications for how the grid must serve these facilities.

A data center receives power from the utility through one or more high-voltage feeds, typically at transmission or sub-transmission voltage — 69 kV, 138 kV, or even 230 kV for the largest facilities. This power passes through utility-owned substations and then through customer-owned switchgear, which distributes it within the facility. Between the utility feed and the servers themselves lies a series of power conditioning and backup systems designed to ensure that the servers never — not for a fraction of a second — lose power. The centerpiece of this architecture is the uninterruptible power supply, or UPS, system: large banks of batteries (historically lead-acid, increasingly lithium-ion) that can instantly assume the load if the utility feed is lost. The UPS provides what the industry calls "ride-through"

capability — the ability to sustain the load for a brief period, typically five to fifteen minutes, while backup generators start and come online. Those backup generators — almost always diesel reciprocating engines, though natural gas turbines and fuel cells are gaining ground — can sustain the facility for hours or even days, limited primarily by fuel supply. The entire system is designed so that a failure at any single point does not result in a loss of power to the IT equipment.

The data center industry has formalized this approach to redundancy through the Uptime Institute's tier classification system, which defines four levels of facility reliability. A Tier I facility has a single path for power and cooling, with no redundancy — suitable for small or non-critical installations. A Tier II facility adds redundant components (an extra UPS module, an extra cooling unit) but retains a single distribution path, designated as N+1 redundancy, meaning that for every N components required to serve the load, one additional component is available as a spare. A Tier III facility introduces multiple distribution paths, so that maintenance can be performed on one path without interrupting the other — the facility is "concurrently maintainable." A Tier IV facility, the highest classification, provides full fault tolerance through 2(N+1) redundancy: two completely independent power and cooling paths, each with its own redundant components, so that the facility can sustain any single failure — or even a simultaneous failure and a planned maintenance event — without interruption. Tier IV facilities can deliver uptime of 99.995 percent or better, which translates to fewer than 26 minutes of downtime per year. Hyperscale data centers operated by the major cloud providers typically meet or exceed Tier III standards, and many critical facilities — those serving financial institutions, government agencies, or the cloud providers' own core infrastructure — are designed to Tier IV.

The metric most commonly used to evaluate data center energy efficiency is Power Usage Effectiveness, or PUE, defined as the ratio of total facility power consumption to the power consumed by the IT equipment itself. A PUE of 2.0 means that for every watt consumed by a server, an additional watt is consumed by cooling, lighting, power conversion losses, and other overhead. The industry average PUE has historically hovered around 1.5 to 1.6, meaning that roughly a third to forty percent of a typical data center's electricity consumption goes to non-computing functions, with cooling representing the dominant share. The hyperscale operators — Google, Microsoft, Amazon — have driven their PUEs down to 1.1 to 1.2 through aggressive engineering: free cooling using outside air in temperate climates, evaporative cooling systems that exploit the latent heat of water evaporation, advanced airflow management that separates hot and cold air streams within the server hall, and, increasingly, direct liquid cooling in which coolant is circulated directly to the hottest components on the server boards. The cooling challenge is intensifying as chip power densities increase — a trend driven by the computational demands of artificial intelligence — and the data center industry's water consumption, which can reach several million gallons per day for a large campus using evaporative cooling, has become a significant environmental and political concern in water-stressed regions.

The critical distinction between data centers and other large industrial loads — and the distinction that makes data centers uniquely challenging from a grid perspective — is the combination of enormous scale, near-zero tolerance for interruption, and complete inflexibility. A large aluminum smelter may draw hundreds of megawatts, but smelters have historically participated in interruptible service arrangements and demand response programs, accepting occasional curtailment in exchange for lower

rates. A data center, by contrast, cannot tolerate even a momentary interruption without risking data loss, service disruption, and, in the case of the major cloud providers, contractual penalties under service level agreements that guarantee uptime to customers worldwide. This is a load that demands the grid's highest quality of service — and that maintains its own redundant generation precisely because it cannot trust the grid alone to provide it.

Load Characteristics

The electrical load profile of a data center is, in a word, flat. Unlike residential load, which peaks in the late afternoon and drops to a trough in the early morning hours, or commercial load, which follows the rhythm of the business day, data center load is essentially constant — twenty-four hours a day, seven days a week, fifty-two weeks a year. Servers do not sleep. Cloud computing workloads are distributed across time zones, so that a facility in Virginia may be serving peak-hour users in Europe while simultaneously serving overnight batch processing jobs in Asia. The result is a load shape that is nearly invariant — a steady, unrelenting draw of megawatts that looks, on a load duration curve, like a horizontal line. This flat profile means that data centers contribute to both peak and baseload demand, but because they cannot be curtailed during peak events, they effectively raise the minimum reserve margin that the system must maintain.

The scale of individual facilities has grown dramatically over the past decade. In the early years of the commercial data center industry — the late 1990s and early 2000s — a large facility might draw ten to twenty megawatts. Today, individual hyperscale data centers routinely draw fifty to one hundred megawatts, and the largest facilities under construction or in planning are designed for two hundred to three hundred megawatts or more. When multiple facilities are co-located in a campus configuration — as is standard practice for the hyperscale operators — the aggregate load can reach one to two gigawatts in a single county, rivaling the output of a large nuclear power plant. This load is concentrated in an extraordinarily small physical footprint: a data center that draws one hundred megawatts may occupy a building of one hundred thousand to two hundred thousand square feet, yielding a power density — measured in watts per square foot — that far exceeds any other building type. A modern office building might draw five to ten watts per square foot; a data center may draw five hundred to one thousand or more. This extreme power density means that serving data centers requires not just generation capacity but high-capacity transmission and distribution infrastructure concentrated in specific locations — infrastructure that, in many cases, does not yet exist.

From the perspective of a grid operator — an RTO like PJM, as described in Chapter 9, or ERCOT, as described in Chapter 7 — data center load is essentially non-dispatchable. The grid operator cannot call upon a data center to reduce its consumption during a capacity emergency in the way that it can dispatch a demand response resource or curtail an interruptible industrial load. Data centers are, in the language of grid operations, firm load of the most inflexible kind. This inflexibility, combined with the sheer magnitude of the load, makes data centers among the most challenging categories of demand that grid operators have ever had to accommodate — a challenge that is compounded by the speed at which

this load is materializing.

* * *

II. The Hyperscale Era: Cloud Computing and the Concentration of Load

The Rise of the Hyperscalers

The data center industry's transformation from a fragmented collection of enterprise server rooms and colocation facilities into a sector dominated by a handful of enormous operators is one of the most consequential developments in the recent history of the American economy — and, as this chapter argues, one of the most consequential developments in the recent history of the American grid. The rise of cloud computing — the model in which businesses and individuals rent computing capacity from centralized providers rather than operating their own servers — has driven an extraordinary concentration of computational infrastructure in the hands of three companies: Amazon Web Services, Microsoft Azure, and Google Cloud Platform. Together, these three "hyperscalers" control approximately two-thirds of the global cloud infrastructure market and operate hundreds of data centers across the United States and around the world.

The economics driving this concentration are straightforward and powerful. Data centers exhibit significant economies of scale in virtually every dimension of their operation: the cost per megawatt of cooling infrastructure declines as facilities grow larger; the cost per unit of networking equipment is lower when amortized across more servers; the operational staff required to maintain a facility does not scale linearly with its size; and the purchasing power of a hyperscale operator — buying servers, storage devices, networking equipment, and electricity in bulk — far exceeds that of a smaller competitor. These economies of scale, combined with the network effects inherent in cloud platforms (developers build applications for the platforms with the most users, which attracts more users, which attracts more developers), have produced a market structure in which the large operators are getting larger while smaller players are increasingly confined to niche roles.

The hyperscalers have adopted a "campus" model of deployment, in which multiple data center buildings are co-located in a single area — often within a few miles of one another — sharing fiber-optic connectivity, power feeds, and support infrastructure. A single campus may comprise five, ten, or even twenty individual buildings, each drawing fifty to two hundred megawatts, yielding an aggregate campus load of one to several gigawatts. This concentration of load in specific geographies is the source of many of the grid challenges examined in this chapter.

Data Center Alley: Northern Virginia as Ground Zero

No place on Earth better illustrates both the promise and the peril of the data center phenomenon than Loudoun County, Virginia, a suburban county in the outer ring of the Washington, D.C. metropolitan area that has become, improbably, the epicenter of the global internet. Loudoun County and the surrounding area of northern Virginia — known in the industry as "Data Center Alley" — host the largest concentration of data centers on the planet: more than three hundred facilities, with dozens more under construction or in planning, collectively drawing multiple gigawatts of power.

The reasons for this concentration are both historical and structural. Northern Virginia's emergence as a data center hub traces to the 1990s, when the Metropolitan Area Ethernet (MAE-East) internet exchange point — one of the original interconnection points for the commercial internet — was established in the area. Proximity to this exchange point, and to the federal government customers (the intelligence community, the Department of Defense, civilian agencies) that represent a large and lucrative market for computing services, attracted the first wave of data center construction. The presence of extensive fiber-optic infrastructure, relatively low electricity costs compared to other East Coast locations, a favorable regulatory and tax environment (Virginia enacted data center tax incentives that have been periodically renewed and expanded), and the availability of land suitable for large-scale development created a self-reinforcing cycle: each new data center made the area more attractive for the next one, as the density of fiber connectivity increased and the local workforce developed specialized expertise.

The utility serving the vast majority of this load is Dominion Energy Virginia, a vertically integrated, investor-owned utility regulated by the Virginia State Corporation Commission. Dominion now faces a challenge unprecedented in the modern history of American electric utilities: a sustained surge in load growth, concentrated in a specific geographic area, that has overwhelmed the planning assumptions, infrastructure capacity, and regulatory frameworks upon which its operations depend. Dominion has reported load growth rates in northern Virginia that are measured not in the one to two percent annual increases that characterized normal utility growth in the twentieth century, but in rates of twenty percent or more per year in certain service territories — growth rates that recall the earliest decades of electrification, when the grid was first being built. The utility has had to construct new substations, new transmission lines, and new distribution infrastructure on an accelerated timeline, often encountering permitting delays, community opposition, and supply chain constraints that extend the time required to complete projects.

The challenge extends beyond Dominion to PJM Interconnection, the regional transmission organization that operates the wholesale electricity market and manages the transmission grid across a thirteen-state territory from Virginia to Illinois, as described in Chapter 9. PJM's interconnection queue — the process through which new generators and large new loads request permission to connect to the grid — has become severely congested, with data center interconnection requests competing for limited transmission capacity alongside the renewable energy projects that are essential to meeting state and federal decarbonization targets. The queue backlog, which extends to several years for some projects, has become a chokepoint that slows both the deployment of clean energy and the construction of new data

centers, creating frustration on all sides.

The phenomenon that began in northern Virginia is now spreading to other regions. Central Ohio — particularly the area around Columbus and New Albany, where Amazon, Google, Microsoft, and Meta have all built or announced major campuses — has emerged as a significant data center market, with load growth that is straining American Electric Power's (AEP) transmission system. Central Texas, particularly the area around San Antonio and the corridor between Austin and Dallas, has attracted data center investment drawn by ERCOT's relatively low wholesale electricity prices and the availability of renewable energy. The Charlotte and Research Triangle regions of North Carolina are seeing major data center buildouts, as is the Phoenix metropolitan area in Arizona — a location that benefits from low land costs and abundant solar energy but faces significant water constraints. The Portland-Seattle corridor in the Pacific Northwest, long attractive for its relatively cool climate and abundant hydroelectric power, continues to draw data center investment. In each of these regions, the pattern is similar: data center load growth that far exceeds historical forecasts, utilities scrambling to expand infrastructure, and communities debating the costs and benefits of hosting these enormous facilities.

The Geographic Logic of Data Center Siting

The geography of data center deployment differs fundamentally from the geography of the loads that the grid was built to serve. For most of the grid's history, electricity demand followed population: load was concentrated in cities and suburbs, in industrial regions and commercial districts, in the places where people lived and worked and made things. Data centers, by contrast, need not be located near the people they serve. A user in New York City streaming a video from Netflix may be served by a data center in Oregon; a financial analyst in Chicago running queries against a cloud database may be accessing servers in Virginia. The speed of light in fiber-optic cable — roughly two hundred kilometers per millisecond — means that for most applications, a data center can be located hundreds or even thousands of miles from its users with negligible impact on performance. Only certain latency-sensitive applications — high-frequency financial trading, real-time gaming, some industrial control systems — require data centers to be located in close physical proximity to their users.

The result is that data center siting decisions are driven by a set of factors that have little to do with proximity to population and everything to do with the economics and availability of infrastructure. The most important factors include the availability and cost of electricity — both the price per kilowatt-hour and the physical capacity of the local transmission and distribution system to deliver large quantities of power; the availability of fiber-optic connectivity — data centers require multiple redundant fiber paths to ensure network reliability; state and local tax incentives — which can amount to hundreds of millions of dollars over the life of a facility; the availability of water for cooling, particularly in regions where evaporative cooling is used; the risk of natural disasters — earthquakes, hurricanes, floods, tornadoes — that could damage the facility or disrupt its power and network connections; and the regulatory environment — both the speed and predictability of permitting processes and the general posture of state and local government toward data center development.

This set of siting criteria creates a geography of load that cuts across the traditional boundaries of utility service territories and wholesale market regions. Data center load materializes in places where utilities did not expect it, in quantities that dwarf historical experience, and on timelines that are measured in months rather than the years or decades that characterize traditional infrastructure planning. The fundamental mismatch between the speed at which data centers can be designed and constructed — a large hyperscale facility can be built in eighteen to twenty-four months — and the speed at which the transmission and generation infrastructure required to serve them can be permitted, financed, and built — typically five to ten years or more for major transmission lines and large power plants — is one of the central challenges that this chapter addresses.

* * *

III. The AI Inflection: A Step Change in Power Demand

From Cloud to AI: A Qualitative Shift

If cloud computing represents the steady, sustained growth of data center demand — a trend that has been underway for two decades and that, by itself, would have eventually forced a reckoning with the grid's capacity — then artificial intelligence represents a step change, a discontinuity that has compressed the timeline and amplified the magnitude of the challenge. The distinction is important because it explains why, after years of gradual growth, data center electricity demand has suddenly become front-page news and a topic of urgent concern for grid planners, utility executives, and policymakers.

Cloud computing workloads — web hosting, email, file storage, software-as-a-service applications, streaming video — are computationally modest on a per-user basis. A single server can serve thousands of users simultaneously for these tasks, and the aggregate power consumption, while enormous when summed across millions of servers, grew in rough proportion to the number of users and the volume of data stored and transmitted. The emergence of large language models and generative artificial intelligence — technologies that reached public awareness with the release of ChatGPT in November 2022 but that had been developing in research laboratories for several years prior — introduced a fundamentally different computational profile. Training a large language model — the process of adjusting billions or trillions of numerical parameters by repeatedly processing vast quantities of text and other data — is among the most computationally intensive tasks ever undertaken. A single training run for a frontier AI model may require tens of thousands of specialized processors (graphics processing units, or GPUs, and purpose-built AI accelerators) operating continuously for weeks or months, consuming tens of megawatts of power throughout. NVIDIA's H100 GPU, the workhorse of AI training in the 2023-2025 period, draws approximately 700 watts at peak; its successor, the B200, draws upward

of 1,000 to 1,200 watts. A training cluster containing 25,000 such chips — a scale that the leading AI companies are now deploying — draws 25 megawatts or more from the GPU chips alone, before accounting for cooling, networking, storage, and other support infrastructure.

AI inference — the process of running a trained model to serve user queries, generate images, write code, or perform other tasks — is less power-intensive on a per-query basis than training, but the aggregate demand is potentially even larger because inference occurs continuously at massive scale as millions of users interact with AI-powered services. Each query to a large language model may consume ten to one hundred times more computational resources — and therefore ten to one hundred times more electricity — than a conventional web search. As AI capabilities are integrated into search engines, productivity software, customer service systems, autonomous vehicles, drug discovery pipelines, and countless other applications, the aggregate inference demand is projected to grow enormously. The combination of training and inference workloads has transformed the data center from a facility that houses servers performing relatively modest computations into a facility that houses what are, in effect, supercomputers operating continuously at full capacity — with commensurate power requirements.

The Load Growth Projections

The magnitude of the projected growth in data center electricity demand is, by any historical standard, extraordinary. In the mid-2020s, total data center electricity consumption in the United States is estimated at approximately 20 to 25 gigawatts of continuous demand — roughly four to five percent of total national electricity consumption. Multiple credible forecasts project this figure growing to 40 to 80 gigawatts or more by 2030, depending on assumptions about the pace of AI adoption, the efficiency trajectory of semiconductor technology, the degree to which AI workloads are concentrated in the United States versus distributed globally, and the extent to which physical infrastructure — power, cooling, fiber, land — can be deployed rapidly enough to keep pace with demand.

To appreciate the scale of this projected growth, consider that the total installed generating capacity of the state of Texas — the largest electricity-consuming state in the nation and the subject of Chapter 7 — is approximately 140 gigawatts, serving a peak demand of roughly 85 gigawatts. Adding 30 to 50 gigawatts of new data center load to the national grid would be roughly equivalent, in electrical terms, to adding an entire additional Texas-sized economy to the system — except that the new load would be concentrated in a handful of specific locations rather than distributed across the state, would demand near-perfect reliability rather than the occasional controlled interruptions that Texas has experienced, and would materialize in a matter of years rather than the century over which Texas's current grid was built. The comparison is imperfect but instructive: it conveys the raw magnitude of the challenge in terms that a reader familiar with the preceding chapters can readily grasp.

The uncertainty range surrounding these projections is enormous and warrants emphasis. The most bullish forecasts — those projecting data center load approaching or exceeding 80 gigawatts by 2030 — assume continued exponential growth in AI adoption, sustained investment in training ever-larger models, and a concentration of this investment in the United States. The more conservative projections

— those in the 35 to 45 gigawatt range — assume that efficiency improvements in AI chips and algorithms partially offset demand growth, that some AI workloads migrate to other countries (particularly those with lower electricity costs or more favorable regulatory environments), and that physical infrastructure constraints — the unavailability of power, cooling water, or suitable sites — slow the pace of deployment. Even the conservative projections, however, represent a rate of load growth that the American grid has not experienced since the mid-twentieth century, and that the current institutional and physical infrastructure is poorly equipped to handle.

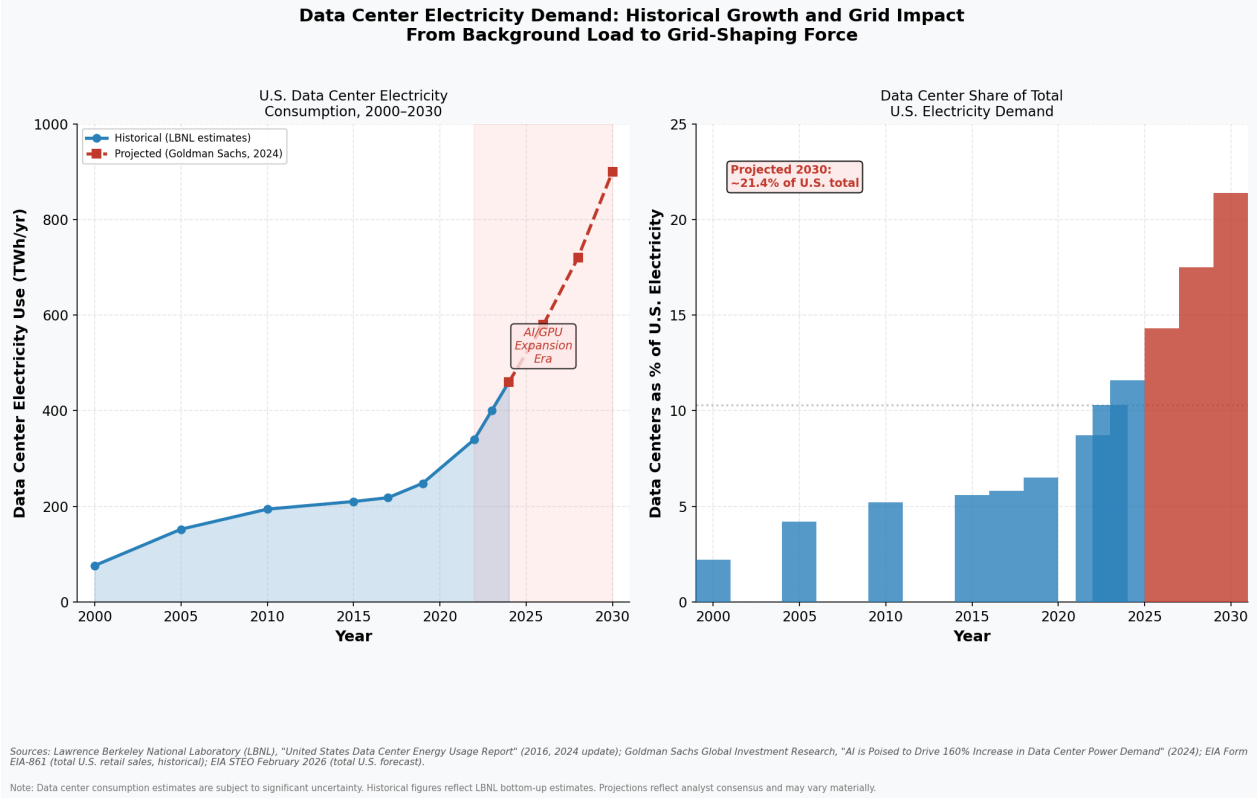

Figure 12.1: Data Center Electricity Demand — Historical Growth and Grid Impact (Source: LBNL, Goldman Sachs, EIA)

The "Load Growth Surprise"

For the better part of two decades, the dominant assumption in American utility planning was that electricity demand growth was effectively over. The efficiency gains driven by LED lighting, high-efficiency appliances, improved building insulation, and the transition from incandescent to electronic displays — collectively producing what energy economists call the "decoupling" of electricity consumption from economic growth — led most utilities and grid planners to project flat or very slowly growing load, on the order of zero to one percent per year. Integrated resource plans filed by utilities with state public utility commissions in the 2010s typically assumed little or no need for new generating capacity beyond what was required to replace retiring coal plants and integrate mandated renewable

energy.

Data centers have shattered this assumption with a speed and thoroughness that has left the utility industry scrambling to respond. Georgia Power, a subsidiary of Southern Company, revised its load growth forecast upward by several gigawatts in a single planning cycle, driven largely by data center and industrial load commitments in the Atlanta metropolitan area. Dominion Energy Virginia, as noted above, has seen load growth in northern Virginia that exceeds anything in its corporate history. Duke Energy, serving the Carolinas, has received data center interconnection requests that would, if fully realized, require gigawatts of new generation and transmission capacity. American Electric Power, serving much of Ohio, Indiana, and West Virginia, has experienced similar surges in its territory. In each case, the utility has had to confront a reality that is deeply uncomfortable: the generation, transmission, and distribution infrastructure that was planned and built for a flat-demand world is insufficient for the world that is now materializing, and the lead times required to build new infrastructure — five to ten years for major transmission projects, seven to twelve years for new nuclear or large gas plants, three to five years for solar and wind farms including interconnection delays — mean that the gap between demand and supply may persist for years.

This planning crisis is compounded by a cruel irony: the same interconnection queue backlogs that delay the construction of data centers also delay the construction of the clean energy resources needed to serve them, as both categories of projects compete for limited transmission capacity and finite engineering and permitting resources. The tension between accommodating data center load growth and meeting clean energy targets — a tension that plays out differently in every region depending on the market structure, regulatory framework, and resource mix — is one of the defining challenges of American energy policy in the 2020s.

* * *

IV. Grid Planning and Infrastructure Impacts

Transmission and Distribution Stress

The physical infrastructure of the grid — the transmission lines, substations, transformers, and distribution feeders described in Chapters 1 and 2 — was designed, built, and sized to serve a particular geography of demand. That geography reflected the distribution of population, commerce, and industry across the American landscape: dense urban cores served by robust transmission and distribution networks; suburban rings with adequate but not excessive capacity; rural areas with long, radial distribution lines sized for low-density load. Data centers disrupt this geography by imposing enormous, concentrated loads in locations that may have been, until recently, lightly loaded suburban or even

semi-rural areas with infrastructure sized accordingly.

A single large data center may require a dedicated feed at transmission voltage — 138 kV or 230 kV — with a capacity of one hundred to three hundred megawatts. A campus of five or ten such facilities requires infrastructure comparable to what would serve a small city. In many cases, this infrastructure does not exist and must be built from scratch: new transmission lines, new substations, new high-voltage switchgear, new distribution transformers. In northern Virginia, Dominion Energy has undertaken a massive program of transmission and substation construction specifically to serve data center load, including new 230 kV lines and the expansion of existing substations to accommodate additional transformer capacity. These projects face the same permitting, land acquisition, and community opposition challenges that confront any major transmission project — challenges discussed at length in Chapter 2 — but with the additional pressure of a customer that expects to energize its facility in eighteen to twenty-four months, not the five to ten years that a major transmission project typically requires.

The interconnection queue problem is particularly acute. As described in Chapter 9's discussion of PJM, the interconnection process — through which new generators and large new loads apply for permission to connect to the transmission grid — has become severely congested. Data center interconnection requests now constitute a significant share of total queue entries at PJM and other RTOs, competing with the massive volume of renewable energy interconnection requests that have accumulated in the wake of the Inflation Reduction Act's incentives (discussed in Chapter 10). FERC has undertaken interconnection queue reform — most notably through Order 2023, issued in 2023 — designed to streamline the process and clear the backlog, but the fundamental tension remains: the queue is a bottleneck because the physical transmission system cannot accommodate all the projects seeking to connect, and expanding the transmission system is a slow, expensive, and politically contested process.

The clustering of data centers in specific areas also creates localized reliability challenges. When multiple large loads are served from the same transmission corridor, the loss of a single line or transformer can affect a disproportionate amount of load. Transmission planners must design the system to withstand contingencies — the loss of any single element without causing cascading failures, as required by NERC reliability standards discussed in Chapter 11 — and the concentration of data center load raises the stakes of these contingency analyses. Moreover, the fact that data centers maintain their own backup generation creates complex interactions during grid disturbances: if a large data center campus simultaneously starts dozens of diesel generators during a grid event, the voltage and frequency transients on the local distribution system can be significant, potentially affecting other customers in the area.

Generation Adequacy

The most fundamental infrastructure challenge posed by data center load growth is the need for new generation. Every additional megawatt of data center demand must be matched by a megawatt of generation capacity — and not just any capacity, but capacity that is available around the clock, every

day, with the high reliability that data center customers demand. This requirement has profound implications for the generation mix and for the trajectory of decarbonization.

Several utilities have responded to data center-driven load growth by delaying or reversing planned retirements of coal and natural gas plants. Georgia Power's decision to defer the retirement of certain coal units, for example, was driven in part by the need to maintain adequate capacity in the face of unexpectedly strong load growth. New natural gas combined-cycle plants are being proposed and constructed in multiple regions specifically to serve data center demand, raising difficult questions about the compatibility of data center growth with state and federal commitments to reduce greenhouse gas emissions. The Inflation Reduction Act's incentives for clean energy, discussed in Chapter 10, create a powerful economic case for renewable generation, but solar and wind resources are intermittent and cannot, without storage, provide the around-the-clock power that data centers require. The missing money problem described in Chapter 6 — the difficulty of ensuring adequate investment in generation capacity in wholesale electricity markets — takes on new dimensions in a world of rapid load growth: more demand should, in theory, raise wholesale prices and improve generator economics, but the speed at which demand is growing may outpace the speed at which new generation can be built, creating reliability risks in the interim.

Perhaps the most striking development in the intersection of data center demand and generation planning is the nascent nuclear renaissance catalyzed by tech company investment. Microsoft's 2023 agreement with Constellation Energy to restart Three Mile Island Unit 1 — the undamaged reactor at the site of the nation's most infamous nuclear accident, a plant that had been shut down in 2019 for economic reasons — exemplifies the scale of ambition. The restarted plant, rated at 835 megawatts, would provide dedicated carbon-free baseload power under a long-term power purchase agreement. Amazon has invested in small modular reactor (SMR) developers, including a partnership with Energy Northwest to site SMRs at an existing nuclear site in Washington state. Google has signed an agreement with Kairos Power, a developer of a novel fluoride-salt-cooled reactor design, for the deployment of advanced reactors to serve Google's data center load. These investments represent a remarkable convergence of interests: the tech companies need large quantities of reliable, carbon-free power that matches their 24/7 load profile, and nuclear energy — uniquely among zero-carbon generation technologies — can provide it. Whether these projects will be completed on time and on budget, given the nuclear industry's troubled construction history in the United States, remains an open question. But the fact that the world's wealthiest and most technically sophisticated companies are willing to commit billions of dollars to nuclear energy is, by itself, a significant development in the energy landscape.

Resource Adequacy and Reliability Implications

The resource adequacy frameworks described in Chapters 5 and 6 — the capacity markets, reserve margin requirements, and planning reserve calculations through which RTOs and utilities ensure that sufficient generation exists to meet projected demand with an acceptable margin of reliability — are being tested by data center load growth in ways that expose both their strengths and their limitations.

Data centers do not merely add megawatts to the system — they add megawatts of a particularly challenging kind. Because data center load is essentially non-curtailable, it cannot participate in demand response programs that provide a critical buffer during capacity emergencies. When a grid operator faces a supply shortfall during a winter storm or a summer heat wave — events discussed at length in Chapter 7's treatment of ERCOT during Winter Storm Uri — it relies on the ability to shed load in a controlled fashion, curtailing interruptible industrial customers, activating demand response contracts, and, as a last resort, implementing rolling blackouts that cycle through residential neighborhoods. Data center load is, for practical purposes, exempt from all of these measures. A cloud provider that loses power to a data center — even for a few seconds — may face contractual penalties, reputational damage, and, in the worst case, data loss or service outages affecting millions of users. Grid operators know this, and they are reluctant to curtail data center load except in the most extreme emergencies.

The practical effect is that data center load increases the reserve margin that the system must maintain to achieve a given level of reliability. If twenty percent of a system's load is non-curtailable (because it consists of data centers and other critical facilities), the grid operator has less flexibility to manage supply shortfalls, and must therefore maintain a larger cushion of excess generation. This higher reserve margin requirement means more generation capacity must be built and maintained — at a cost that is ultimately borne by all ratepayers, raising the equity concerns discussed later in this chapter. Grid operators face a paradox identified earlier: load growth should theoretically ameliorate the missing money problem by increasing demand and therefore prices, but the inflexibility of the growing load makes the system harder — not easier — to manage during the stress events that most severely test reliability.

* * *

V. Power Procurement and the Clean Energy Imperative

Corporate Clean Energy Commitments

The major technology companies that are the largest consumers of data center electricity have, without exception, made ambitious public commitments to power their operations with clean energy. Google has pledged to operate on 24/7 carbon-free energy by 2030 — not merely matching its annual electricity consumption with renewable energy certificates, but ensuring that every kilowatt-hour consumed by every Google facility is matched, in real time, by a kilowatt-hour of carbon-free generation. Microsoft has committed to becoming carbon negative by 2030, meaning that the company aims to remove more carbon dioxide from the atmosphere than it emits, including the emissions associated with its entire supply chain. Amazon has pledged to power its operations with one hundred percent renewable energy

by 2025 and to achieve net-zero carbon emissions by 2040.

These commitments have made the technology sector the largest corporate buyer of renewable energy in the world, driving an enormous volume of power purchase agreements (PPAs) with wind, solar, and battery storage developers. A PPA is a long-term contract — typically ten to twenty years — under which a corporate buyer agrees to purchase the output of a renewable energy project at a fixed price, providing the revenue certainty that the project developer needs to secure financing and proceed with construction. Tech company PPAs have financed gigawatts of new renewable capacity across the United States, and the scale of these commitments continues to grow.

The question of whether these PPAs actually reduce emissions — the question of "additionality" — is more complex than it might appear. If a tech company signs a PPA for the output of a solar farm that would have been built regardless — perhaps because it is located in a state with a renewable portfolio standard that creates sufficient demand for solar generation — then the PPA does not cause any additional clean energy to be deployed. The company may claim credit for the clean energy, but total emissions are unchanged. A PPA is genuinely "additional" only if it causes a clean energy project to be built that would not otherwise have been built — a counterfactual that is inherently difficult to establish. The most rigorous approaches to corporate clean energy procurement attempt to maximize additionality by financing projects in regions where the grid is heavily reliant on fossil fuels, or by structuring contracts that go beyond simple energy purchases to include requirements for firm, dispatchable clean energy (such as renewable-plus-storage combinations or nuclear power).

The evolution from annual to hourly matching represents a significant tightening of the standard. Under the traditional approach, a company might purchase enough renewable energy certificates to match its total annual electricity consumption, regardless of whether the renewable generation occurred at the same time as the consumption. Under hourly matching — the approach pioneered by Google and increasingly adopted by other companies — the goal is to ensure that clean energy generation matches consumption on an hour-by-hour (or even finer) basis. This is a far more demanding standard, because solar generation peaks during the day while data center load is constant, requiring the company to procure clean energy from a diverse portfolio of resources — solar, wind, geothermal, nuclear, hydroelectric, battery storage — that collectively cover all hours of the day and night. The pursuit of 24/7 carbon-free energy is one of the most powerful market signals in the contemporary energy landscape, and it is reshaping investment decisions across the generation sector.

The Nuclear Option

The scale, reliability requirements, and flat load profile of data center demand make nuclear power uniquely attractive as a power source — a convergence of interests that is producing the most significant investment in nuclear energy in the United States in decades, even as the nuclear industry's commercial prospects had appeared, until recently, to be in terminal decline.

Nuclear power plants produce electricity continuously, at full capacity, with capacity factors typically exceeding ninety percent — a performance profile that almost perfectly matches the flat, 24/7

load shape of a data center. Nuclear generation produces no direct greenhouse gas emissions, making it compatible with the clean energy commitments of the major technology companies. And nuclear plants are long-lived assets — a well-maintained plant can operate for sixty years or more — providing a stability of supply that appeals to companies making multi-decade infrastructure investments.

The Three Mile Island restart is perhaps the most symbolically potent example of this convergence. Unit 1 of the Three Mile Island nuclear plant — a pressurized water reactor rated at 835 megawatts, not to be confused with Unit 2, which suffered a partial meltdown in 1979 in the most serious accident in the history of the American nuclear industry — had operated safely and reliably for decades before being shut down in 2019 because its revenues from PJM's wholesale electricity market were insufficient to cover its operating costs. Microsoft's agreement with Constellation Energy to purchase the plant's output under a long-term PPA at a price that makes restart economically viable represents a remarkable second act for a facility that had become a symbol of nuclear energy's troubled history in the United States.

The technology companies' investments in advanced reactor designs — small modular reactors, molten salt reactors, high-temperature gas-cooled reactors — reflect a longer-term bet that next-generation nuclear technology can be deployed at data center sites, providing dedicated, on-site, carbon-free power. Whether this vision is realized depends on a host of technical, regulatory, and economic factors that are beyond the scope of this chapter, but the direction of investment is unmistakable.

Behind-the-Meter and On-Site Generation

The imperative for reliable, carbon-free power, combined with the delays and frustrations of the utility interconnection process, has led some data center operators to explore the possibility of generating their own power on-site — a practice known as "behind-the-meter" generation because the generating equipment is located on the customer's side of the utility meter, rather than on the grid. On-site generation options include natural gas turbines, fuel cells (both hydrogen and natural gas-fueled), and, prospectively, small nuclear reactors.

Behind-the-meter generation raises a set of questions that go to the heart of the regulatory compact described in Chapters 3 and 4. If a large data center generates its own power on-site, it may reduce or eliminate its purchases from the utility — but it still relies on the grid for backup power during maintenance periods or equipment failures, and it still benefits from the grid's role in maintaining system stability and reliability. The costs of the grid — the transmission lines, substations, distribution infrastructure, and system operations that maintain reliability for all users — are typically recovered through volumetric charges (charges per kilowatt-hour consumed). If a large customer reduces its consumption by generating its own power, it contributes less to the recovery of these fixed costs, which must then be spread across the remaining customers. This phenomenon — sometimes called "grid defection" or "cost shifting" — is a concern that extends beyond data centers to include rooftop solar, industrial cogeneration, and any other form of distributed generation, but the enormous scale of data center loads gives it particular urgency. A single data center campus that goes behind the meter could

shift millions of dollars per year in grid costs to other ratepayers.

Regulators and utilities are grappling with how to address this challenge. Some approaches include standby charges (fees paid by behind-the-meter generators for the option of drawing on the grid when needed), demand charges (fees based on peak power draw rather than total energy consumption), and interconnection fees that reflect the cost of maintaining grid infrastructure regardless of how much energy the customer consumes. The design of these rate structures has significant implications for the economics of both data centers and the grid, and for the equitable allocation of costs across customer classes.

* * *

VI. Policy, Regulation, and Community Impact

Economic Development vs. Grid Stress

Data centers are, by any measure, significant economic investments. A single hyperscale data center facility may represent five hundred million to over one billion dollars in capital investment, including the building, the power and cooling infrastructure, and the computing equipment housed within. A multi-building campus may represent several billion dollars. This investment generates property tax revenue that can be transformative for the host jurisdiction — Loudoun County, Virginia, derives a substantial share of its property tax revenue from data center facilities, enabling it to maintain relatively low tax rates for residential property owners while funding high-quality public services.

Yet data centers are unusual among large industrial investments in the number of permanent jobs they create — or, more precisely, in the number they do not create. A data center that costs one billion dollars to build and draws two hundred megawatts of power may employ only fifty to one hundred permanent operations and maintenance staff once construction is complete. The construction phase generates hundreds or thousands of temporary jobs, but the operational workforce is small by the standards of traditional industrial facilities of comparable capital cost. A manufacturing plant of equivalent investment might employ a thousand or more workers. This low employment density means that the economic development case for data centers rests primarily on property tax revenue and construction activity rather than on sustained job creation — a distinction that has become increasingly salient as communities weigh the costs and benefits of hosting these facilities.

Those costs are not negligible. Data centers generate significant noise — the constant hum of cooling fans and backup generators, sometimes audible from considerable distances, has prompted complaints and regulatory responses in some jurisdictions. Their water consumption, as noted earlier, can be substantial, particularly in facilities that use evaporative cooling, and this consumption has become a

flashpoint in water-stressed regions. Their visual impact — large, windowless, industrial-style buildings, often surrounded by fencing and security infrastructure — has drawn objections from residents accustomed to the suburban or rural character of their communities. And the strain they place on local infrastructure — roads, utilities, emergency services — can exceed the capacity of jurisdictions that were not designed to accommodate large industrial loads.

These concerns have produced a growing backlash in some communities. Several jurisdictions in Virginia have enacted moratoriums on new data center construction, seeking time to evaluate the cumulative impacts of the facilities already built and to develop more comprehensive planning frameworks. The phenomenon is not unique to the United States: Ireland, the Netherlands, and Singapore have all imposed restrictions on new data center construction or energy consumption, driven by concerns about grid capacity, water consumption, or greenhouse gas emissions.

Regulatory Challenges

The regulatory frameworks described in Chapters 3 through 6 — the system of state public utility commissions, federal regulation by FERC, and wholesale market oversight by RTOs and ISOs — were designed for a world in which load growth was gradual and geographically distributed, and in which the principal categories of customers — residential, commercial, and industrial — had well-understood demand characteristics. Data centers challenge these frameworks in multiple dimensions.

State public utility commissions face novel questions about rate design and cost allocation. Should data centers pay the same rates as other large commercial or industrial customers, or should they be placed in a separate rate class that reflects their unique characteristics — their flat load profile, their extreme reliability requirements, their concentrated geographic impact? Some argue that data centers should pay a premium for their inflexibility — a reflection of the additional reserve margin and infrastructure costs they impose on the system. Others argue that the flat, predictable load profile of data centers makes them, in some respects, an ideal customer — easy to forecast, consistent in their demand, and potentially useful as anchor tenants that provide a stable revenue base for new generation and transmission investments.

FERC's role is primarily in the areas of interconnection policy, transmission planning, and wholesale market design. As discussed above, interconnection queue reform is a pressing concern, and FERC has taken steps through Order 2023 and subsequent proceedings to streamline the queue process and reduce the backlog. Transmission planning — the process through which RTOs and transmission owners identify the need for new transmission facilities and allocate their costs — must increasingly account for data center load growth, which may require new approaches to long-term demand forecasting and infrastructure investment. The interaction of data center demand with wholesale market design is also significant: the entry of large, inelastic loads into wholesale markets affects price formation, congestion patterns, and the economics of generation investment in ways that existing market models may not fully capture.

The water-energy nexus deserves particular attention. Data centers that use evaporative cooling can

consume several million gallons of water per day for a large campus — a volume that is significant in any context but especially so in the water-stressed regions of the American West and South where data center development is expanding. The permitting and allocation of water resources is governed by a patchwork of state laws, water rights doctrines, and local regulations that were not designed with data center water consumption in mind. The competition between data centers and other water users — agriculture, municipal supply, ecosystem needs — is a source of growing tension that will intensify as data center deployments expand into water-constrained regions.

The Equity Question

The most fundamental policy question raised by data center load growth may be the simplest to state and the most difficult to resolve: who should pay for the grid infrastructure required to serve data centers? The traditional approach — in which the costs of generation, transmission, and distribution infrastructure are socialized across all ratepayers in proportion to their consumption — means that residential customers, small businesses, and other commercial and industrial users all contribute to the cost of infrastructure built primarily to serve data centers. In regions where data center load growth is driving the need for billions of dollars in new transmission and generation investment, the resulting rate increases can be significant — and they fall disproportionately on customers who derive no direct benefit from the data center facilities.

This concern has prompted some states and utilities to explore alternative cost allocation mechanisms. Contribution-in-aid-of-construction (CIAC) requirements, under which a new large customer pays some or all of the cost of the infrastructure required to serve it, are one approach. Special rate classes for data centers, with rate structures designed to reflect the full cost of service including the system-level impacts of their demand, are another. Network upgrade charges, in which interconnecting customers pay for the transmission system upgrades required to accommodate their load, are a third. Each approach has advantages and disadvantages, and the appropriate choice depends on the specific circumstances of the utility, the state regulatory framework, and the competitive dynamics of the data center market — since states that impose high costs on data centers may find that developers choose to build in more favorable jurisdictions instead.

The equity question is not merely economic but also, in a deeper sense, about the purpose of the grid. The regulated utility model described in Chapters 3 and 4 was premised on a social compact: in exchange for a guaranteed monopoly and the opportunity to earn a fair rate of return, the utility was obligated to serve all customers in its territory at just and reasonable rates. This compact assumed that the grid existed to serve people — households, businesses, and communities. Data centers strain this compact because they represent an enormous demand for grid services that primarily benefits the shareholders and customers of technology companies, many of whom are located far from the communities where the infrastructure is built and the costs are borne. The question of how to update the social compact of regulated electricity to accommodate this new reality is one of the most important and least resolved questions in American energy policy.

* * *

Conclusion: The Grid as Digital Infrastructure

Data centers represent a fundamental shift in what the American electric grid is for. For the better part of a century, the grid existed to serve people — to light homes, power factories, cool offices, and run the machines that underpin daily life. The load it carried was, in a deep sense, human-shaped: it followed the rhythms of human activity, concentrating in the places where people lived and worked, rising in the morning and falling at night, peaking in summer when air conditioners labored and ebbing in spring when the weather was mild. Data centers serve machines — servers performing computation that may be located hundreds or thousands of miles from the humans who ultimately benefit from it. A query typed into a search engine in San Francisco may be processed by a server in Iowa; a video streamed to a phone in Boston may originate from a data center in Oregon. This decoupling of load from population creates a new geography of electricity demand — one that concentrates enormous quantities of load in places chosen for fiber connectivity, tax incentives, and power availability rather than proximity to people — and challenges every assumption embedded in the grid's physical, institutional, and regulatory architecture.

The challenge is acute precisely because data center load combines the worst characteristics from a grid planning perspective. It is massive in scale — individual facilities drawing hundreds of megawatts, campuses aggregating gigawatts. It is concentrated in specific locations — Data Center Alley, central Ohio, central Texas — overwhelming infrastructure designed for more distributed loads. It demands near-perfect reliability — 99.999 percent uptime, with zero tolerance for even momentary interruptions. It is essentially non-curtailable — grid operators cannot shed data center load during emergencies the way they can curtail interruptible industrial customers or activate demand response programs. And it is growing faster than any category of load in living memory — at rates that have turned decades of flat-demand utility planning assumptions upside down. The American grid was not designed for this. Adapting it will require investment in generation, transmission, and distribution infrastructure on an enormous scale; innovation in market design, rate structures, and planning methodologies; and institutional reform in the regulatory frameworks that govern the allocation of costs, benefits, and risks across the diverse stakeholders of the electric power system. The magnitude of the undertaking is comparable, in some respects, to the original electrification of the American economy — a process that took decades, consumed vast resources, and remade the physical and institutional landscape of the nation.

Yet the story of digital load and the grid is not solely one of stress and strain. The next chapter examines a very different kind of digital electricity consumer — Bitcoin mining — which, in sharp contrast to the data center loads described here, offers a model of how computation can serve not merely as a consumer of grid services but as a provider of them. Where data centers are inflexible, Bitcoin mining is interruptible; where data centers demand perfect reliability, Bitcoin mining can tolerate

frequent and prolonged curtailment; where data centers concentrate load in ways that stress the grid, Bitcoin mining can locate at points on the grid where excess generation would otherwise go to waste. The contrast between these two forms of digital load — one rigid, one flexible; one a burden on the grid, the other potentially a balm — illuminates fundamental questions about the relationship between electricity consumption, economic value, and the architecture of the power system. That contrast is the subject of Chapter 13.

* * *

Key Concepts: data center, hyperscale, cloud computing, artificial intelligence, power usage effectiveness (PUE), uninterruptible power supply (UPS), N+1 redundancy, Tier IV reliability, load growth, interconnection queue, Data Center Alley, behind-the-meter generation, power purchase agreement (PPA), additionality, small modular reactor (SMR), grid defection, water-energy nexus, rate design, contribution-in-aid-of-construction

Chapter 13: Bitcoin Mining and the Grid — Digital Load as Grid Infrastructure

Introduction: Beyond the Energy Headline

Few topics in contemporary energy discourse generate as much heat — and as little light — as Bitcoin mining's relationship with the electric power grid. Since the cryptocurrency's rise to mainstream awareness in the mid-2010s, its energy consumption has become one of the most frequently cited statistics in debates about technology, sustainability, and the future of the global energy system. Headlines declaring that Bitcoin "uses more electricity than Argentina" or "could single-handedly derail the Paris Agreement" have become a staple of popular science journalism, and the question of whether proof-of-work mining constitutes an unconscionable waste of energy has animated editorial pages, congressional hearings, and environmental advocacy campaigns alike. The intensity of this discourse is understandable — in an era of climate crisis, any large and rapidly growing source of electricity demand invites scrutiny, and the abstract, intangible nature of Bitcoin's output makes it easy to frame the enterprise as pure consumption with no corresponding social benefit.

This chapter does not attempt to resolve the normative question of whether Bitcoin mining is a worthwhile use of electricity. That question, while important, belongs to a different kind of book. What this chapter does attempt — consistent with the analytical framework applied throughout the preceding chapters — is to examine Bitcoin mining as an engineering and economic phenomenon within the American power system. From the perspective of a grid operator managing real-time supply-demand balance, a transmission planner evaluating congestion patterns, a renewable energy developer struggling with curtailment and negative pricing, or a state regulator weighing economic development against environmental impact, Bitcoin mining presents a genuinely novel category of electric load — one whose physical and economic characteristics have measurable, and in some cases operationally significant, consequences for how the grid functions.

The contrast with the data center loads examined in Chapter 12 could hardly be more striking. Where conventional data centers demand near-perfect reliability and cannot tolerate even brief

interruptions, Bitcoin mining is almost infinitely interruptible — each successive hash computation is statistically independent of the last, a property known as memorylessness, and nothing adverse happens if a mining facility runs at sixty percent, twenty percent, or zero percent capacity for minutes or hours. Where data centers must locate near population centers and fiber-optic interconnection points, mining requires nothing from its location except electricity and a basic internet connection. Where data center load is rigid, inflexible, and essentially non-dispatchable from the grid operator's perspective, mining load can be dialed up or down on a second-by-second basis in response to price signals or grid operator commands. These properties — interruptibility, location-agnosticism, price-elasticity, modularity, and a constant load factor — are precisely the properties that grid operators have long wished more loads possessed, particularly as the generation mix shifts toward variable renewable resources whose output does not follow demand.

This chapter proceeds in six parts. It begins by characterizing the electrical profile of Bitcoin mining facilities, explaining what makes them operationally distinct from other large industrial loads. It then examines mining's role as a demand response resource, with particular attention to ERCOT's pioneering experience in Texas — the jurisdiction that has, as discussed in Chapter 7, become the de facto laboratory for grid innovation in the United States. The chapter next considers mining's capacity to monetize stranded and curtailed energy, a function with significant implications for renewable energy deployment and transmission planning. A section on flare gas mitigation explores one of the most environmentally consequential applications of mining technology. The chapter then analyzes how co-located mining operations affect the project economics of renewable energy development. Finally, a survey of the evolving regulatory landscape situates mining within the broader policy debates that will shape the American grid in the decades ahead.

* * *

I. The Electrical Profile of a Bitcoin Mining Facility

The Load Characteristics That Matter

To understand how Bitcoin mining interacts with the grid, one must first understand what a mining facility actually looks like from the perspective of a utility engineer or grid operator. The popular image of Bitcoin mining — rows of blinking computers in a nondescript warehouse — captures the visual reality but obscures the electrical characteristics that matter for grid operations. What distinguishes mining load from other forms of large-scale electricity consumption is not its magnitude, though that is substantial, but rather a constellation of physical and economic properties that together create a load profile unlike anything the grid has previously encountered.

The first and perhaps most consequential of these properties is location-agnosticism. Unlike a hospital, which must be located near the population it serves; unlike a manufacturing plant, which must be situated along supply chains and near labor markets; unlike a conventional data center, which must minimize latency to end users by locating near population centers and fiber-optic interconnection points — a Bitcoin mining facility requires nothing from its location except access to electricity and a basic internet connection. The computational work of mining — the repeated calculation of SHA-256 hash functions in a probabilistic search for valid blocks — is not latency-sensitive. A mining machine in rural West Texas competes on precisely equal footing with one in downtown Houston or suburban Virginia. The data transmitted to and from the Bitcoin network is minimal — measured in kilobytes per second, not the gigabytes per second that a cloud computing data center must handle. This means that mining operations can locate wherever electricity is cheapest, a property that has profound implications for how mining interacts with the geography of the grid and the spatial distribution of generation and transmission resources. As Senator Ted Cruz observed at the Texas Blockchain Summit, "the beauty of Bitcoin mining is that if you can connect to the internet, you can use that energy and derive value from those renewables in a way that would be impossible otherwise."

The second critical property is interruptibility. A Bitcoin mining facility can reduce its electricity consumption from full load to zero within seconds — literally at the flip of a switch, or more precisely, at the execution of a software command. Unlike an aluminum smelter, where sudden power interruption can cause molten metal to solidify in the reduction pots, destroying equipment worth hundreds of millions of dollars and requiring weeks of restart procedures, a mining facility experiences no physical damage whatsoever from instantaneous shutdown. The machines simply stop computing, and they resume computing the moment power is restored. There is no work-in-progress that is lost, no product that is spoiled, no chemical reaction that must be carefully managed through a shutdown sequence. The economic cost of curtailment is simply the foregone mining revenue for the hours during which the machines are idle — a cost that is predictable, quantifiable, and, as we shall see, often more than offset by the payments miners receive for providing this interruptibility as a service to the grid.

The third property is a constant, predictable load factor. When operating, mining machines draw power at a constant rate — the load does not fluctuate with production cycles, shift changes, weather patterns, or seasonal demand as industrial loads typically do. A facility drawing 100 megawatts will draw 100 megawatts continuously, twenty-four hours a day, seven days a week, three hundred and sixty-five days a year, unless deliberately curtailed. This constant baseload profile is, from a grid planning perspective, highly predictable and easy to model — a welcome contrast to the variable, weather-dependent generation profiles of wind and solar resources that are rapidly growing as a share of the generation mix.

The fourth property is extreme price-elasticity. Because mining is a commodity business in which the revenue per kilowatt-hour of electricity consumed is determined by the Bitcoin network's difficulty adjustment and the market price of Bitcoin — variables over which individual miners have no control — the profitability of a mining operation is overwhelmingly determined by its electricity cost. This creates a powerful economic incentive to consume only the cheapest available electricity. Miners are, in effect, natural arbitrageurs of the electricity market — they are economically programmed to seek out and

consume energy at the lowest available price, which in practice means energy that is surplus, stranded, off-peak, or otherwise undervalued by the market. This price-seeking behavior has consequences for both where miners locate and when they choose to operate, consequences that interact in complex ways with wholesale electricity market dynamics. Critically, miners engage in what economists call "economic dispatch" — they react to real-time prices and simply do not run their equipment if electricity prices get too high. During the February 2021 winter storm in Texas, even the highest-end equipment (Bitmain Antminer S19s) had an economic turn-off point of approximately $480 per megawatt-hour — well below the $9,000 per megawatt-hour price cap that ERCOT's market reached during the crisis. The grid does not need to rely on the beneficence of miners to expect them to curtail during scarcity; as profit-maximizing entities, they have a clear and powerful economic motive to do so.

The fifth property is modularity and portability. Modern mining operations are frequently deployed in standardized shipping containers — self-contained units that include mining hardware, power distribution, cooling systems, and network connectivity. These containerized units can be transported by truck, deployed on a concrete pad with an electrical interconnection, and brought online within days. They can also be relocated when economic conditions change — when a power purchase agreement expires, when local electricity prices rise, or when a more attractive opportunity emerges elsewhere. This modularity means that mining load can be deployed at a scale and in locations that match the characteristics of available power, rather than requiring power infrastructure to be built to serve the load, as is typically the case with conventional industrial development. Mining is, in the language of one industry analysis, "strongly fractionalizable" — a single shipping container of miners can viably exploit a sub-one-megawatt source of energy, something no aluminum smelter or chemical plant could contemplate.

Taken together, these five properties — location-agnosticism, interruptibility, constant load factor, price-elasticity, and modularity — describe a category of load that is, from the grid operator's perspective, almost ideally suited to several of the operational challenges that have intensified as the American power system transitions toward a higher penetration of variable renewable generation. No other major category of electric load possesses all five of these properties simultaneously, and it is this combination that makes Bitcoin mining's interaction with the grid worthy of the detailed examination that follows. Prior to Bitcoin, there simply was no industrial load resource that satisfied all of these qualities with such fidelity. Some industrial consumers of load possessed some of them — aluminum smelting, for instance, has a degree of location agnosticism, as has been well-documented with Alcoa smelters co-locating with abundant energy resources. Certain types of factories like aluminum arc furnaces and paper mills have a degree of interruptibility, but only for short periods and with significant latency. None could provide the flexibility, fractional attenuation, or response times that Bitcoin miners can.

Scale and Growth

The scale of Bitcoin mining in the United States has grown dramatically since 2020, driven by a

combination of factors: China's ban on cryptocurrency mining in mid-2021, which displaced enormous amounts of hash rate to jurisdictions with cheap electricity and favorable regulatory environments; the availability of inexpensive natural gas and renewable energy in several American regions; and a generally permissive regulatory environment in key states. By the mid-2020s, estimates of total Bitcoin mining load in the United States ranged from approximately five to more than ten gigawatts — a figure that, while uncertain due to the fragmented and sometimes opaque nature of the industry, places mining in the same order of magnitude as major industrial sectors. For comparison, the entire aluminum smelting industry in the United States consumes approximately three to four gigawatts.

This mining load is geographically concentrated in regions that reflect the price-seeking behavior described above. Texas, with its deregulated wholesale market, abundant wind and solar generation, and welcoming political environment, hosts the largest concentration — estimated at two to four gigawatts, with individual facilities ranging from tens of megawatts to several hundred megawatts. Upstate New York, with its legacy of abundant and inexpensive hydroelectric power from the Niagara and St. Lawrence projects, hosts a significant cluster, though its growth has been constrained by a state-level moratorium on fossil-fuel-powered mining enacted in 2022. The Dakotas and Wyoming, with excellent wind resources and sparse population to compete for electricity, have attracted substantial mining development, as have portions of Appalachia — particularly western Pennsylvania and Kentucky — where access to inexpensive natural gas from the Marcellus and Utica shale formations provides a competitive electricity cost. The Pacific Northwest, with its abundant hydroelectric resources and historically low electricity prices, has also attracted mining operations, though capacity constraints at public utilities have sometimes led to moratoriums on new interconnections.

* * *

II. Demand Response and Controllable Load

The Grid Operator's Problem: Load That Can Think

The fundamental operational challenge of the electric power grid, as established in the opening chapters of this book, is the requirement to maintain continuous, real-time balance between electricity generation and electricity consumption. Because electricity cannot be economically stored at scale — though battery storage is beginning to change this at the margin — every watt consumed must be simultaneously produced, and any imbalance between supply and demand manifests as a deviation in system frequency from its nominal 60 Hz. As discussed in Chapter 1, even small frequency deviations can cause equipment damage, trigger protective relay operations, and, in extreme cases, cascade into widespread blackouts. Grid operators maintain this balance through a combination of dispatchable generation, operating

reserves, and demand response — programs through which large consumers agree to reduce their electricity consumption during periods of grid stress in exchange for compensation.

The value of demand response — load that can be rapidly curtailed at the grid operator's direction — has grown significantly as the generation mix has shifted toward variable renewable resources. When the wind stops blowing or clouds cover solar panels, the resulting drop in generation must be offset either by ramping up dispatchable generators or by reducing demand. As the penetration of inverter-based resources increases and the fleet of dispatchable fossil generators shrinks — trends documented extensively in Chapter 10 — the value of controllable, interruptible load on the demand side has increased commensurately. Texas leads all states in installed wind generation capacity and is expected to continue doubling its renewable capacity over the coming years. Being an islanded grid with a significant and growing portion of energy supply coming from renewables requires ERCOT to procure and utilize more responsive demand response products — with requirements to respond in seconds or even at sub-second frequency — in addition to the more traditional ten-to-thirty-minute response times that most demand response programs have historically required.

The challenge is that most large industrial loads are not easily interruptible. Chemical plants, refineries, steel mills, semiconductor fabrication facilities, and the conventional data centers discussed in Chapter 12 — all of these require continuous power for process or reliability reasons, and even those that can participate in demand response programs typically do so only with significant advance notice, limited duration, and a restricted number of curtailment events per year. The electric load that grid operators have long wished for — load that can curtail instantly, for any duration, at any time, with no advance notice, and with no physical consequences — has historically been a theoretical ideal rather than a practical reality. Bitcoin mining comes closer to this ideal than any other large-scale electric load in the history of the grid.

Bitcoin Miners as Large Flexible Loads in ERCOT

Nowhere has the integration of Bitcoin mining into grid operations been more thoroughly developed than in ERCOT, the independent system operator that manages the Texas grid. As discussed in detail in Chapter 7, ERCOT operates an energy-only market without a capacity market — a design that relies on scarcity pricing to incentivize both the construction of new generation and the provision of demand-side flexibility. This market structure creates naturally strong price signals during periods of grid stress: when supply is tight, wholesale prices can spike to the market cap of $5,000 per megawatt-hour (since reduced from $9,000) — more than one hundred times the typical price. These extreme price signals create powerful economic incentives for flexible loads to curtail during scarcity events, and Bitcoin miners, with their unique interruptibility characteristics, have proven to be among the most responsive participants in ERCOT's demand-side programs.

ERCOT classifies Bitcoin mining facilities as Large Flexible Loads (LFLs), a designation that recognizes their distinctive operational characteristics and subjects them to specific interconnection and operational requirements. As of the mid-2020s, several gigawatts of mining load had registered with

ERCOT under this classification, making mining one of the largest categories of controllable load on the Texas grid.

But the most operationally significant classification is what ERCOT calls a Controllable Load Resource, or CLR — a category of demand response that goes far beyond the traditional ten-to-thirty-minute response windows. A CLR must be able to curtail seventy percent of its targeted load reduction within sixteen seconds of receiving a signal from the grid operator. Before Bitcoin mining, no load type had ever qualified as a CLR in ERCOT. The Houston-based technology company Lancium became, in 2020, the first entity to qualify a load as a CLR — approximately one hundred megawatts of Bitcoin mining load that could be optimized on both a daily and hourly basis to mine when it was economic to do so and to curtail when it was not.

The concept of a CLR is best understood as a power generator in reverse. Instead of adding expensive power to the grid during a period of scarcity, the CLR receives a real-time price signal from the grid operator. If the price exceeds the mining operation's economic turn-off point, the CLR automatically "dispatches down" — curtails consumption — to make way for other, more critical loads. The capacity not reserved as grid insurance is offered into ERCOT's security-constrained economic dispatch (SCED) and will automatically dispatch down when the real-time price exceeds the turn-off point for the mining load. This is a profound innovation: for the first time, instead of relying solely on flexible thermal generation from coal or natural gas plants to meet peak demand, ERCOT has a demand-side resource that can provide equivalent services on a second-by-second basis.

An added benefit to ERCOT in having Bitcoin mining loads as CLRs is that during local shortages or system emergencies, ERCOT can directly curtail the load. For the mining operator, this arrangement is economically attractive: they can sell ancillary services — a bundle of products that give the grid operator the right to curtail the facility's consumption — and collect a premium for doing so, while mining the rest of the time. This premium effectively lowers their all-in power cost, even if they are never actually called upon to curtail. In contrast, a generation resource that sells ancillary services has a real opportunity cost: it must operate below its maximum output in order to retain slack in case it is called upon to increase its power. The CLR model elegantly eliminates this asymmetry.

Miners also participate in ERCOT's Emergency Response Service (ERS) program, which pays participating loads for the contractual commitment to curtail during grid emergencies. When ERCOT issues a curtailment order, participating miners must reduce their consumption within the specified timeframe and maintain that reduction for the duration of the emergency. The economics are compelling: during the relatively few hours per year when ERCOT experiences genuine scarcity conditions, the combination of curtailment payments and avoided exposure to extremely high wholesale prices far exceeds the mining revenue that would have been earned during those hours.

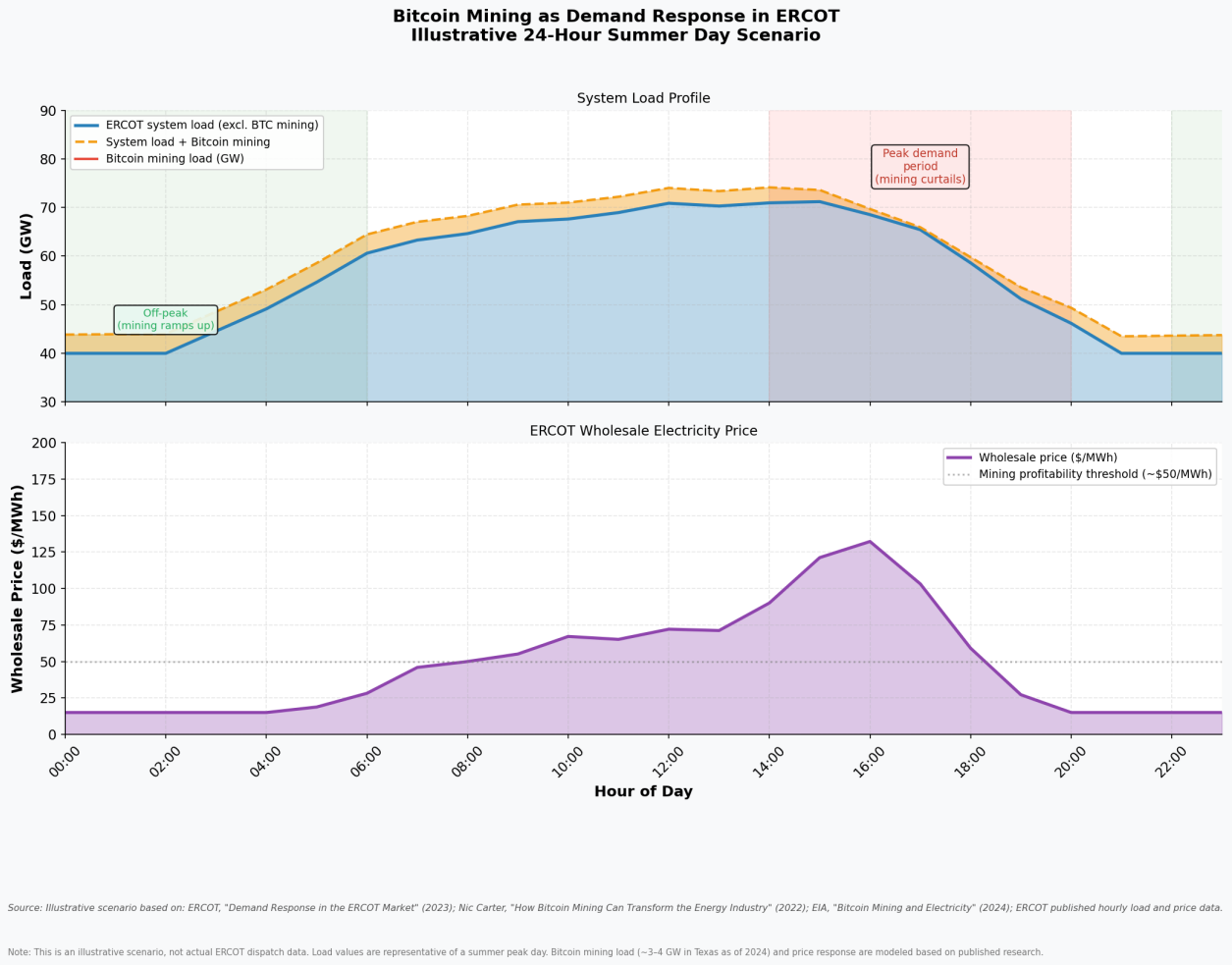

Figure 13.1: Bitcoin Mining as Demand Response in ERCOT — Illustrative 24-Hour Summer Day Scenario (Source: ERCOT, EIA, Nic Carter)

The real-world performance of miners as demand response resources has been documented by both ERCOT officials and independent researchers. During the extreme summer heatwaves that have stressed the Texas grid in recent years — events during which temperatures exceeded 110 degrees Fahrenheit across much of the state, driving air conditioning load to record levels — Bitcoin miners curtailed hundreds of megawatts within minutes of ERCOT's conservation signals. Similarly, during winter storm events that evoked the traumatic memory of Winter Storm Uri in February 2021, miners again curtailed rapidly and reliably. Brad Jones, who served as interim CEO of ERCOT during a critical period following the Uri crisis, has subsequently co-authored research with Nic Carter, a prominent cryptocurrency researcher at Castle Island Ventures, documenting the operational role that mining loads have played in improving grid reliability. Their paper, "Leveraging Bitcoin Miners as Flexible Load Resources for Power System Stability and Efficiency," provides a comprehensive technical analysis, drawing on operational data and economic modeling to demonstrate that well-integrated mining load can function as a de facto operating reserve, reducing the probability of involuntary load shedding during grid emergencies.

The uptime economics make the logic inescapable. Analysis of ERCOT pricing data has shown that in 2021 — the year of Winter Storm Uri — if a power consumer strategically avoided the highest-priced periods and reduced its uptime expectation from one hundred percent to ninety-five percent, its average annual electricity cost fell from $178 per megawatt-hour to a mere $25 per megawatt-hour. The right tail of the price distribution — the five percent of hours during which prices spiked above $100 per megawatt-hour — accounted for the overwhelming majority of total annual electricity costs. Miners, as profit-maximizing entities with the unique ability to tolerate downtime, naturally avoid precisely those hours. Every megawatt of voluntary curtailment by a miner during a scarcity event is a megawatt that does not need to be involuntarily shed from a hospital, a residential neighborhood, or a water treatment plant.

The Net Effect: More Generation, More Flexibility

The presence of mining load on the grid has a dual effect that deserves particular emphasis. On one side, mining provides a consistent, baseload source of demand that eagerly absorbs cheap or negatively-priced power — everything on the left side of the price distribution curve. This improves the economics of energy producers who, for the first time, have a new buyer for their electricity beyond the inflexible grid. Improved generator economics promote the construction of more energy infrastructure and improve the prospects for existing installations. On the other side, as a highly interruptible load that can tolerate downtime, mining ensures that more power is available for households and hospitals during periods of scarcity, when supply trips offline through weather or other disruptions.

Research from Dr. Joshua Rhodes and Dr. Thomas Deetjen at IdeaSmiths LLC has quantified this dual effect. Their analysis of the ERCOT grid found that "operating data centers in a flexible manner can result in a net reduction of carbon emissions" and can "increase the resilience of the grid by reducing demand during high stress times on the grid." Under a scenario in which five gigawatts of flexible data center load was added to the Texas grid with uptime ranging between eighty-five and eighty-seven percent, the flexible load consumed approximately 35.5 million megawatt-hours — but supported the deployment of an additional 39.5 million megawatt-hours of wind and solar energy. In simple terms, the incremental megawatt-hour output from renewable sources exceeded the incremental megawatt-hour consumption from the flexible load. The net effect was carbon negative.

This finding has profound implications for the policy debate around Bitcoin mining and the grid. It suggests that under the right conditions — when mining load is genuinely flexible, when it participates in demand response programs, and when it is located in regions with abundant renewable energy resources — the presence of mining on the grid can actually accelerate decarbonization rather than impede it.

Demand Response Beyond Texas

While Texas has been the most prominent venue for mining's integration into grid demand response, similar dynamics are emerging in other jurisdictions. In upstate New York, where substantial mining

operations have been established near hydroelectric facilities, NYISO has explored the inclusion of mining loads in its demand response programs, though the regulatory environment has been complicated by the state's 2022 moratorium on new fossil-fuel-powered mining facilities. In MISO's territory — spanning the upper Midwest and portions of the South — mining operations in the Dakotas and other wind-rich states have participated in utility-level demand response programs. SPP, whose territory overlaps significantly with the nation's best wind resources in Oklahoma and Kansas, has similarly seen interest from miners seeking to co-locate with wind generation and participate in demand response. The common thread across these jurisdictions is that grid operators, regardless of their market structure, recognize the operational value of large, flexible, controllable loads — and that Bitcoin mining currently represents the largest and most responsive source of such load available. On a percentage of peak demand basis, ERCOT actually lags peers like MISO when it comes to enrolling loads in demand response — suggesting significant untapped potential for flexible mining loads in other markets.

* * *

III. Monetizing Stranded and Curtailed Energy

The Problem of Nonrival Energy

One of the most intellectually compelling arguments for Bitcoin mining's grid-level significance centers on a concept that Nic Carter, drawing on economic theory, has termed "nonrival energy" — energy that is physically produced but that cannot, for practical reasons, reach any willing consumer. This is not a hypothetical phenomenon. It is a routine, large-scale feature of the American power system, and it represents one of the most significant sources of economic waste in the electricity sector. As Carter explains, nonrival energy is "energy that is stranded or curtailed because it cannot reach a population center" — and Bitcoin acts as a unique buyer for it.

The most common cause of nonrival energy is transmission congestion. As discussed throughout this book — and particularly in the analysis of locational marginal pricing in Chapter 5 and the transmission constraints facing renewable development in Chapter 10 — the American grid's transmission infrastructure was largely designed and built to serve a generation fleet of centralized fossil fuel and nuclear plants located near population centers. The rapid build-out of wind generation in the Great Plains and solar generation in the Southwest has created a fundamental mismatch between where electricity is produced and where transmission capacity exists to deliver it. When a wind farm in West Texas generates electricity at three o'clock in the morning — a time of abundant wind and minimal demand — and the transmission lines connecting West Texas to the load centers of Dallas, Houston, and San Antonio are already at capacity, that electricity has no path to any consumer. The grid operator must

either curtail the wind farm or accept the congestion and allow locational marginal prices at the wind farm's node to drop to zero or below.

The real-time dynamics of this phenomenon are visible in ERCOT market data. During a representative period in early October in West Texas, wind and solar generation averaged over twenty gigawatts, with the power lines connecting West Texas to the major load centers in the east operating at maximum capacity. With so much wind and solar online, West Texas power prices averaged just three dollars per megawatt-hour across three days, with several hours settling at negative prices. When the wind dropped — from twenty gigawatts at midnight to just two gigawatts by noon — power flowing across the West Texas transmission lines also dropped, causing West Texas prices to spike to parity with the rest of ERCOT. A grid with even higher wind and solar penetration would face these problems in abundance.

This curtailment is not trivial in scale. ERCOT alone curtailed more than six percent of total wind generation in some recent years, representing billions of kilowatt-hours of clean energy that was available but could not be delivered to consumers. In California, the curtailment problem for solar energy has been even more severe during spring months, when abundant sunshine coincides with mild temperatures and low air conditioning demand — the phenomenon that produces the famous duck curve discussed in Chapters 8 and 10. CAISO has curtailed billions of kilowatt-hours of solar energy annually, and the problem is projected to intensify as solar capacity continues to grow faster than either transmission capacity or energy storage deployment.

Beyond transmission congestion, nonrival energy arises from geographic isolation. There are, as Senator Cruz has noted, "a lot of places on earth where the sun shines a lot and the wind blows a lot but there aren't any power lines. And so it's not economically feasible to use that energy." Hydroelectric dams in remote mountain valleys, run-of-river installations on distant waterways, wind farms sited in locations selected for their excellent wind resources rather than their proximity to transmission infrastructure — all of these can produce energy that exceeds the capacity of available transmission to export.

Bitcoin as the Buyer of Last Resort

Bitcoin mining's location-agnosticism enables it to function as what might be called a "buyer of last resort" for nonrival energy. Because mining operations require nothing from their location except electricity and basic internet connectivity, they can be sited at the exact geographic point where surplus energy exists — at the base of a wind farm in the Texas Panhandle, beside a hydroelectric facility in rural New York, adjacent to a solar installation in the Arizona desert. By consuming energy that would otherwise be curtailed, the mining operation converts energy with a market value of zero into economic output.

The economic significance extends beyond the mining operation itself. When a wind farm operator knows that a co-located mining facility will purchase all output that cannot be sold into the wholesale market, the effective capacity factor of the wind farm increases — more of its potential energy

production generates revenue, and the financial risk of curtailment is reduced. This improved capacity factor translates directly into better project economics: higher revenue, faster debt repayment, improved return on equity for investors, and a lower levelized cost of energy that makes the project more competitive in power purchase agreement negotiations. In this way, mining demand at the point of generation functions as a subsidy to renewable energy development — not a subsidy paid by taxpayers or ratepayers, but one paid by the Bitcoin network's block reward.

The Arcane Research report of 2022 — published by the Norwegian cryptocurrency research firm K33, with a foreword by Nic Carter — documented this dynamic in considerable detail, examining case studies of co-located mining and wind generation. The report found that mining's function as a flexible, co-located load significantly improved the economics of wind projects in transmission-constrained areas, and that the carbon intensity of the electricity consumed by co-located miners was substantially lower than the average grid intensity.

In the hydroelectric context, the dynamic is even more straightforward. Upstate New York's legacy hydroelectric facilities — descendants of the same Niagara Falls installations that powered the early alternating current grid, as discussed in Chapter 1 — produce clean, reliable, baseload electricity at very low marginal cost. But the transmission infrastructure connecting these remote facilities to the downstate load centers of New York City and its suburbs has been constrained for decades. Mining operations that locate near these hydroelectric facilities consume electricity that is already being produced, that is carbon-free, and that would otherwise be either curtailed or sold at depressed local prices that undervalue the resource. Similar dynamics play out at hydroelectric facilities in the Pacific Northwest and, across the international border, at Hydro-Quebec's vast installations.

Transmission Congestion as Opportunity

The relationship between mining and transmission congestion deserves particular emphasis because it connects to one of the most persistent challenges facing the American grid. As documented in Chapter 10, the United States faces a multi-decade backlog of proposed transmission projects, with interconnection queues stretching to thousands of projects and many years of wait time. Bitcoin mining does not solve the transmission problem — that requires the policy and engineering solutions discussed in Chapter 10 — but it can mitigate the economic waste that the transmission bottleneck produces in the interim. By locating at the constrained end of a transmission corridor and consuming electricity that cannot be exported, mining reduces curtailment and provides revenue that sustains generators' financial viability while the transmission system catches up. Having a significant quantity of flexible loads on the grid to dial down consumption during rapid drawdowns in renewable generation also helps attenuate spikes in power prices, without requiring as much support from less efficient combustion turbine peakers.

* * *

IV. Flare Gas Mitigation and Methane Reduction

The Flaring Problem

Among the most environmentally significant applications of Bitcoin mining technology — and one of the least appreciated in the public discourse — is the capture and productive use of natural gas that would otherwise be flared or vented at oil production sites. When crude oil is extracted from the earth, it is frequently accompanied by "associated gas" that rises to the surface through the same wellbore. In mature oil-producing regions with well-developed pipeline infrastructure, this gas is captured and transported to processing facilities. But in newer or more remote production areas — particularly in the Permian Basin of West Texas and New Mexico, and the Bakken Formation of North Dakota — pipeline infrastructure has not kept pace with the explosive growth of oil production.

The scale of flaring in the United States is enormous. As Senator Cruz has noted, "fifty percent of the natural gas in this country that is flared is being flared in the Permian right now in West Texas." That energy is, in Cruz's words, "just being wasted — being wasted because there is no transmission equipment to get that natural gas where it could be used the way natural gas would ordinarily be employed; it's just being burned." The World Bank's Global Gas Flaring Reduction Partnership has identified the United States as one of the world's largest flarers, and satellite imagery from NOAA and NASA reveals the Permian Basin and Bakken Formation as bright clusters of flare light visible from space. Beyond the direct waste of a valuable energy resource, flaring produces significant air quality impacts in nearby communities and contributes measurably to greenhouse gas emissions. Moreover, open flaring is imperfect combustion — studies have consistently found that flares achieve combustion efficiencies well below one hundred percent, meaning that a portion of the gas passes through the flame unburned, entering the atmosphere as methane, a greenhouse gas with approximately eighty times the global warming potential of carbon dioxide over a twenty-year horizon.

Wellhead Mining: Converting Waste to Computation

Companies such as Crusoe Energy, Giga Energy, and several smaller operators have developed portable, containerized systems that deploy directly at wellheads in oil-producing regions. These systems consist of a small natural gas generator — typically a reciprocating engine or microturbine — that burns the waste gas to produce electricity, which powers containerized Bitcoin mining hardware. The entire system is modular, truck-portable, and can be operational within days of arriving at a well site.

The environmental logic is compelling. Combustion in an enclosed generator is significantly more complete than combustion in an open flare. While a flare might achieve ninety to ninety-five percent combustion efficiency under ideal conditions — and substantially less in windy conditions or with wet gas — a properly maintained generator achieves ninety-nine percent or higher combustion efficiency.

This means that for every unit of gas processed, the generator releases substantially less unburned methane to the atmosphere than the flare it replaces. The reduction in methane emissions represents a meaningful net decrease in greenhouse gas impact, given methane's dramatically higher warming potential.

The economic logic is equally compelling. For the oil producer, the waste gas represents not merely a zero-value byproduct but an actual liability — flaring requires permits, incurs regulatory scrutiny, and increasingly carries penalties or royalty obligations. A mining operator who arrives at the wellhead and offers to take the gas at no cost — or pays a modest price for it — transforms a liability into a revenue source. For the mining operator, the electricity produced from waste gas is extraordinarily inexpensive — often below two cents per kilowatt-hour. The result is a commercial arrangement that benefits both parties and produces a measurable environmental improvement over the status quo.

By the mid-2020s, several hundred megawatts of mining capacity were operating on flare gas across the Permian Basin, Bakken Formation, and other oil-producing regions. Several environmental organizations, including the Environmental Defense Fund, have cautiously acknowledged that wellhead mining represents a form of harm reduction — not an ideal solution to the flaring problem, but an improvement over the status quo that has the practical advantage of being deployable immediately without requiring new pipeline infrastructure or regulatory mandates.

* * *

V. Renewable Energy Project Economics

The Revenue Stack Problem

The economics of renewable energy development are shaped by a fundamental tension that has intensified as wind and solar penetration has grown. The capital costs of these technologies have declined dramatically, but the revenue that renewable projects earn is increasingly undermined by the very success of the technology: as more wind and solar capacity is added to the grid, the wholesale electricity prices during periods of high renewable output are depressed — often to zero or below — reducing the revenue that each project earns. This is the duck curve problem documented in Chapters 8 and 10, and it represents one of the most significant barriers to continued renewable energy deployment.

Mining as a Supplemental Revenue Stream

Co-located Bitcoin mining addresses this problem by providing a guaranteed purchaser for electricity during periods when wholesale market prices are lowest. When the wholesale price drops below the

miner's breakeven electricity cost, the miner purchases the output directly from the renewable generator at a negotiated price that, while low, is higher than the zero or negative wholesale price the generator would otherwise receive. When wholesale prices rise above the miner's breakeven — during peak demand periods or scarcity events — the miner curtails, freeing the electricity to be sold into the wholesale market at the higher price. The result is a revenue profile that is smoother and more predictable than the volatile wholesale market alone would provide.

This arrangement functions as what some analysts have termed an "economic battery" — a mechanism that stores the economic value of cheap, off-peak electricity in a non-physical form. A chemical battery stores energy by converting it to electrochemical potential; a co-located mining operation stores the economic value in the form of Bitcoin, which can be sold at any time and whose value is independent of the temporal and spatial dynamics of the electricity market. This analogy is imperfect — a real battery feeds stored energy back to the grid during periods of high demand — but it captures the economic function that mining serves for the generator's revenue stack.

The implications for project finance are significant. Renewable energy projects are capital-intensive, with the vast majority of lifetime costs incurred at construction. The availability of a co-located mining offtake agreement can improve a project's financial profile by reducing revenue variance, establishing a floor on expected revenue, and providing a creditworthy counterparty for a portion of the project's output. Several renewable energy developers have begun incorporating mining offtake into their project design from the outset — sizing their generation capacity larger than the available transmission interconnection can support, with the excess output designated for co-located mining consumption. This approach allows the developer to capture the full energy potential of an excellent wind or solar site without being constrained by available transmission capacity, effectively using mining as a substitute for the transmission infrastructure that cannot be built quickly enough.

* * *

VI. Policy, Regulation, and Controversy

The Energy Consumption Critique

Any serious analysis must engage honestly with the critique that mining consumes an enormous amount of electricity. The Cambridge Centre for Alternative Finance (CCAF) at the University of Cambridge maintains the most widely cited estimate, the Cambridge Bitcoin Electricity Consumption Index, which in the mid-2020s placed annual global consumption in the range of one hundred to one hundred and fifty terawatt-hours. The United States accounts for a substantial share — perhaps thirty to sixty terawatt-hours annually. Critics contend that this electricity could serve other purposes and that mining's

existence increases total demand, which at the margin is met by fossil-fuel generation. Lyn Alden, in her widely-cited analysis "Bitcoin's Energy Usage Isn't a Problem. Here's Why," has provided a rigorous, math-heavy breakdown that contextualizes Bitcoin's energy use globally and explains how stranded hydroelectric power and flared gas fit into the economic equation.

The Counterargument: Marginal vs. Average Emissions

The response, developed most comprehensively by Nic Carter and Lyn Alden, hinges on a distinction between average and marginal emissions intensity. The average emissions intensity of the American grid — total carbon dioxide divided by total electricity produced — is a useful metric for many purposes, but it does not accurately describe the emissions impact of a specific load that consumes electricity at specific times and in specific locations. What matters is the marginal emissions intensity — the emissions produced by the specific generators that increase their output in response to mining demand.

Because miners are economically incentivized to consume the cheapest available electricity — surplus renewable energy, off-peak nuclear or hydro baseload, stranded energy in transmission-constrained areas — the marginal emissions intensity of mining load is often substantially lower than the average grid intensity. A miner consuming curtailed wind energy in West Texas at three in the morning imposes zero marginal emissions, because that energy would have been produced regardless. Conversely, a miner consuming electricity from a natural gas peaker during tight supply would impose high marginal emissions — but this is precisely the scenario in which the miner's economic incentives lead it to curtail. The "Bitcoin Net Zero" paper by Carter and Ross Stevens of NYDIG formalized this argument with empirical analysis, estimating that the marginal emissions intensity of Bitcoin mining is significantly lower than the average grid intensity.

Regulatory Responses

Texas has adopted a relatively welcoming approach, enacting legislation that provides tax incentives for mining facilities that participate in demand response programs while establishing registration requirements for large operations. The Texas framework ties incentives to demand response participation, ensuring that mining facilities are integrated into grid management rather than operating as purely passive consumers.

New York has taken a starkly different approach. In November 2022, Governor Kathy Hochul signed legislation imposing a two-year moratorium on new and renewed permits for fossil-fuel-powered mining operations — the first state-level restriction of its kind. The moratorium does not apply to mining powered by renewable energy, implicitly acknowledging the environmental difference between the two categories.

At the federal level, regulatory attention has been episodic. The Department of Energy's Energy Information Administration undertook a data collection effort that faced legal challenges from the industry. Congressional hearings have examined mining's energy impact with sharply divergent

assessments from witnesses. The academic paper "Valorization of Curtailed Power: Enhancing Grid Flexibility Through Cryptocurrency Mining," published in the Journal of Climate Change Research, has been cited in policy debates as evidence that the binary framing of mining as either "good" or "bad" for the environment is inadequate.

The Grid Operator's Perspective

The entities most directly responsible for managing the physical grid have generally taken a more pragmatic view than either mining's critics or its advocates. ERCOT officials have consistently stated that mining load, when properly integrated into demand response programs, is operationally beneficial. The grid operators' perspective is shaped by their institutional mandate: they are responsible for reliability and economic efficiency, not for making value judgments about the social utility of the loads they serve. From this perspective, a load that curtails instantly during emergencies, that provides predictable baseload demand during normal operations, and that locates where surplus energy exists is — whatever one thinks of its ultimate purpose — a grid asset rather than a grid liability. As the grid becomes increasingly renewable, moving from steady fossil-fuel-powered baseload to more volatile wind and solar power, these kinds of controllable loads will become increasingly critical.

* * *

Conclusion: A New Category of Load

The history of the American electric grid, as traced across the preceding chapters, is in large part a history of the system's adaptation to new categories of demand. Thomas Edison's Pearl Street Station was built to serve incandescent light bulbs. The electrification of American industry transformed the grid into an industrial power system. The post-war suburban boom brought millions of air conditioning units, reshaping the demand profile. The rise of the internet produced the hyperscale data center — a massive, concentrated, reliability-demanding load examined in Chapter 12 that is itself reshaping transmission planning and generation investment.

Bitcoin mining is the latest entry in this succession — and in many ways the most unusual, because it is the first major category of load whose defining characteristic is its willingness to be curtailed. The four properties identified by industry analysts — interruptibility, attenuation (the ability to dial down fractionally rather than simply on or off), unconstrained location agnosticism, and scale independence — are unique in combination. Among industrial load centers, these qualities have no precedent. Prior to Bitcoin, there was no load resource that satisfied all four qualities with such fidelity.

None of this is to suggest that Bitcoin mining is an unalloyed good for the grid or for society. The energy consumption is real, the scale is large, and the ultimate social value of the Bitcoin network is a

question on which reasonable people disagree profoundly. What this chapter has demonstrated is that the interaction between Bitcoin mining and the American power grid is a complex, multidimensional phenomenon that resists simple narratives. The grid does not care about the purpose of the load it serves — it cares about the load's physical characteristics, its responsiveness to price signals, its willingness to curtail during emergencies, and its effect on the economics of the generation fleet. By those operationally relevant criteria, Bitcoin mining is a phenomenon that the architects, operators, and regulators of the American power system must understand — and that, when properly integrated, can contribute to the system's reliability, efficiency, and decarbonization in ways that the headline energy consumption figures alone would never suggest.

The ongoing research by Carter, Jones, and their collaborators represents exactly the kind of rigorous, technically grounded analysis that this new category of load demands. As the American grid continues its historic transformation — integrating hundreds of gigawatts of variable renewable generation, managing the electrification of transportation and buildings, and confronting the cybersecurity and resilience challenges documented in Chapter 11 — the role of flexible, controllable demand will only grow in importance. Bitcoin mining has demonstrated that such demand exists at scale and that it can be integrated into grid operations in ways that benefit the system as a whole. That demonstration, regardless of one's views on cryptocurrency, is a contribution to the engineering and economics of the American power system that warrants serious scholarly attention.

Key Concepts: Bitcoin mining, proof-of-work, hash rate, interruptible load, demand response, Large Flexible Loads (LFL), Controllable Load Resource (CLR), Emergency Response Service (ERS), curtailed energy, nonrival energy, flare gas mitigation, methane reduction, co-location, buyer of last resort, marginal emissions intensity, load flexibility, controllable load, revenue stack, transmission congestion, wellhead mining, economic dispatch, attenuation, scale independence

* * *

Appendix A: Comparative Overview of Major RTOs and ISOs

The following reference summarizes the key characteristics of the seven major Regional Transmission Organizations and Independent System Operators operating in the contiguous United States.

Feature	PJM Interconnection	MISO	ERCOT	CAISO	NYISO	ISO-NE	SPP
Geography	13 states and D.C. (Mid-Atlantic to Midwest)	15 states and Manitoba (Upper Midwest to Gulf Coast)	~90% of Texas	Most of California and part of Nevada	New York State	6 New England states	14 states (Great Plains and Central U.S.)
Population Served	~65 million	~45 million	~27 million	~30 million	~19 million	~14.5 million	~18 million
Interconnection	Eastern	Eastern	ERCOT (islanded)	Western	Eastern	Eastern	Eastern
Market Type	Capacity + Energy	Capacity + Energy	Energy-Only	Energy + RA Program	Capacity + Energy	Capacity + Energy	Energy + RA
Capacity Mechanism	Reliability Pricing Model (RPM) — forward auction 3 years ahead	Planning Resource Auction (PRA) — seasonal	None (ORDC scarcity pricing; Performance Credit Mechanism under development)	Resource Adequacy program administered by CPUC; no centralized capacity market	Installed Capacity (ICAP) market — monthly/seasonal/annual strips	Forward Capacity Market (FCM) — 3-year forward with Pay-for-Performance	Resource adequacy requirements; no centralized capacity auction
FERC Jurisdiction	Yes	Yes	No (regulated by PUCT)	Yes	Yes	Yes	Yes
Distinguishing Feature	Largest wholesale market by volume; robust capacity market	Largest geographic footprint; Multi-Value Project transmission portfolio	Jurisdictional independence; energy-only design; 2021 Winter Storm Uri	Highest solar penetration; duck curve management; Western Energy Imbalance Market	Zone J (NYC) load pocket; Indian Point retirement; CLCPA mandates	Winter fuel security challenge; LNG dependency; Pay-for-Performance penalties	Record wind penetration (>90% in single hours); Integrated Marketplace
Key Policy Challenge	Capacity market design amid state clean energy mandates	Interregional transmission (MISO South seam); resource adequacy tightening	Reliability without capacity payments; weatherization after Uri	Integrating variable renewables; EDAM expansion across the West	Decarbonizing Zone J; replacing Indian Point; 2040 zero-emission target	Winter gas-electric coordination; offshore wind integration	Wind curtailment during low-load periods; westward market expansion

Note: Population figures and market characteristics reflect conditions as of the mid-2020s. Market designs continue to evolve through regulatory proceedings at FERC and state commissions.

Appendix B: Glossary of Key Terms

Alternating Current (AC): Electric current that periodically reverses direction, oscillating at a fixed frequency (60 Hz in North America, 50 Hz in most of the rest of the world). The standard form of electricity on the power grid, enabling efficient voltage transformation through transformers.

Ancillary Services: Services necessary to support the reliable operation of the transmission system, including frequency regulation, spinning reserves, non-spinning reserves, voltage support, and black-start capability. Procured through RTO-administered markets or bilateral contracts.

Area Control Error (ACE): A real-time measurement of a Balancing Authority's deviation from its scheduled net interchange with neighboring areas, adjusted for frequency. ACE is the primary control signal used by automatic generation control systems to maintain supply-demand balance.

Automatic Generation Control (AGC): A centralized control system that automatically adjusts the output of participating generators every few seconds to maintain system frequency and correct Area Control Error.

Balancing Authority (BA): An entity responsible for maintaining the real-time balance between electricity supply and demand within a defined geographic area, and for managing interchange with neighboring Balancing Authorities. There are approximately 60 BAs in the contiguous United States.

Baseload Generation: Generation resources designed to operate continuously at or near full output for extended periods, characterized by high capital costs and low variable (fuel) costs. Historically includes nuclear power plants, large coal-fired steam plants, and run-of-river hydroelectric facilities.

Black Start: The process of restoring a power system from a total blackout without relying on external power. Requires generators with the capability to start independently (black-start resources) and a carefully sequenced restoration plan.

Bulk Electric System (BES): The interconnected electrical generation, transmission, and associated equipment, generally operated at voltages of 100 kV or higher. The system to which NERC reliability standards apply.

Capacity Factor: The ratio of a generator's actual energy output over a period to the maximum output it could have produced operating at full rated capacity for the entire period. Expressed as a percentage; e.g., a nuclear plant operating at 90% capacity factor produces 90% of its theoretical maximum annual output.

Capacity Market: A market mechanism that compensates resources for their commitment to be

available to produce electricity during future periods, separate from payments for energy actually produced. Designed to ensure long-term resource adequacy. Examples include PJM's RPM and ISO-NE's FCM.

Combined-Cycle Gas Turbine (CCGT/NGCC): A power plant that combines a gas turbine (Brayton cycle) with a heat recovery steam generator and steam turbine (Rankine cycle) to achieve higher thermal efficiency (45–54%) than either cycle alone.

Congestion: A condition in which the desired flow of electricity between two points on the transmission network is limited by the physical capacity of the transmission facilities or by operating security constraints. Congestion is reflected in differences in Locational Marginal Prices between constrained locations.

Cost of New Entry (CONE): The annualized cost of building and operating a new generation resource, typically a natural gas combustion turbine, used as a benchmark in capacity market design for setting demand curves and evaluating price outcomes.

Cost-of-Service Regulation: The traditional regulatory model in which a utility's rates are set to recover its prudently incurred costs of providing service plus a regulated rate of return on invested capital (the rate base).

Controllable Load Resource (CLR): In ERCOT, a category of demand response in which a load resource can be dispatched down by the grid operator on a second-by-second basis in response to real-time price signals. CLRs must curtail seventy percent of their targeted load reduction within sixteen seconds. Bitcoin mining facilities were the first loads to qualify as CLRs in ERCOT.

Data Center: A facility housing computing equipment (servers, storage, networking) that provides cloud computing, data storage, and digital services. Data centers are characterized by flat 24/7 load profiles, extreme reliability requirements (99.999% uptime), and high power density. They represent the fastest-growing category of electric load in the United States.

Demand Response (DR): Programs and mechanisms that incentivize or require electricity consumers to reduce or shift their electricity consumption in response to price signals, grid reliability needs, or direct utility/RTO dispatch instructions.

Distributed Energy Resource (DER): A small-scale electricity generation, storage, or demand-side resource located on the distribution system or behind a customer's meter. Includes rooftop solar, home batteries, electric vehicles, smart thermostats, and other devices.

Distribution System: The portion of the electric power system that delivers electricity from substations to end-use customers, typically operating at voltages below 69 kV. Includes distribution feeders, transformers, meters, and associated equipment.

Duck Curve: A graph of net load (total demand minus variable renewable generation) that exhibits a deep midday depression (caused by high solar output) followed by a steep evening ramp (as solar declines and demand increases), named for its duck-like shape. First prominently identified by CAISO.

Electric Reliability Organization (ERO): An organization certified by FERC to develop and enforce mandatory reliability standards for the Bulk Electric System. The North American Electric Reliability Corporation (NERC) is the sole certified ERO.

Electromagnetic Pulse (EMP): An intense burst of electromagnetic energy, potentially produced by a high-altitude nuclear detonation, that can damage or destroy electronic equipment across a wide geographic area.

Energy-Only Market: A wholesale market design in which generators are compensated only for the energy they produce, without separate capacity payments. Relies on scarcity pricing during tight supply conditions to provide investment signals. ERCOT is the primary example in the United States.

Federal Energy Regulatory Commission (FERC): An independent federal agency that regulates the interstate transmission and wholesale sale of electricity, natural gas, and oil. FERC approves RTO tariffs, market rules, and transmission rates, and enforces compliance with market rules.

Forward Capacity Market (FCM): A capacity market design that procures capacity commitments several years in advance of the delivery period. ISO New England's FCM procures capacity three years forward.

Generation: The process of producing electricity from primary energy sources (fossil fuels, nuclear fission, wind, solar radiation, falling water, etc.) at power plants.

Geomagnetic Disturbance (GMD): A disruption of the Earth's magnetic field caused by solar coronal mass ejections, which can induce damaging currents in long conductors including high-voltage transmission lines and transformer windings.

Grid-Following Inverter: A power electronic inverter that synchronizes its output to the voltage and frequency of the grid by measuring and tracking the grid's waveform. Requires an externally established voltage reference (typically from synchronous machines) to operate.

Grid-Forming Inverter: A power electronic inverter that establishes its own voltage and frequency reference, enabling it to operate without relying on synchronous machines for a reference signal. Can provide synthetic inertia, voltage regulation, and black-start capability.

Heat Rate: A measure of the thermal efficiency of a generator, expressed as the number of British thermal units (BTU) of fuel required to produce one kilowatt-hour of electricity. Lower heat rates indicate higher efficiency.

Independent Market Monitor (IMM): An entity, internal or external to an RTO/ISO, responsible for monitoring wholesale market outcomes for evidence of market power abuse, market manipulation, or design flaws.

Independent Power Producer (IPP): A non-utility entity that owns and operates electric generating facilities and sells power at wholesale, typically in organized RTO/ISO markets or through bilateral contracts.

Independent System Operator (ISO): An independent organization that coordinates, controls, and monitors the operation of the electric power system within a defined region. Functionally similar to an RTO, though the terms are sometimes used to reflect differences in organizational history or scope.

Interconnection: One of the three major synchronized alternating current networks in North America: the Eastern Interconnection, the Western Interconnection, and the ERCOT Interconnection. All generators within an interconnection operate in precise synchrony. Interconnections are linked to one another only through asynchronous DC ties.

Inverter: A power electronic device that converts direct current (DC) to alternating current (AC). Used to connect solar panels, batteries, and certain wind turbines to the AC grid.

Inverter-Based Resource (IBR): A generation or storage resource that connects to the grid through a power electronic inverter rather than through a directly coupled synchronous machine. Includes solar photovoltaics, battery energy storage systems, and Type 3 and Type 4 wind turbines.

Large Flexible Load (LFL): In ERCOT, a designation for large electricity consumers — including Bitcoin mining facilities — that can rapidly adjust their consumption in response to grid conditions or price signals. LFLs are subject to specific interconnection and operational requirements.

Locational Marginal Price (LMP): The marginal cost of serving the next increment of load at a specific location on the transmission network, as determined by the RTO's security-constrained economic dispatch algorithm. LMP has three components: the cost of generation (the energy component), the cost of transmission losses, and the cost of transmission congestion.

Load Shedding: The deliberate reduction of electric load, either through voluntary demand response or involuntary disconnection of customers (rolling blackouts), implemented to maintain grid frequency and prevent system collapse during supply shortages.

Merit Order: The ranking of available generation resources from lowest to highest marginal cost of production. The dispatch algorithm "stacks" resources in merit order, dispatching the cheapest resources first to serve load.

Minimum Offer Price Rule (MOPR): A capacity market rule that sets a minimum price at which new or subsidized resources may offer into capacity auctions, intended to prevent the suppression of capacity prices by resources receiving out-of-market revenues (such as state renewable energy subsidies).

Missing Money Problem: The theory that energy market revenues alone are insufficient to support the investment in generation capacity needed to maintain system reliability, because energy prices are subject to offer caps and other mitigation measures that prevent them from rising to the level that would fully compensate peaking resources. The missing money problem is the primary justification for capacity markets.

N-1 Criterion: A reliability planning standard requiring that the electric system be able to withstand the loss of any single element (generator, transmission line, or transformer) without violating operating limits or losing customer load.

Natural Monopoly: A market condition in which a single firm can serve the entire market at lower cost than two or more firms, due to economies of scale and the high fixed costs of the infrastructure required. Electricity transmission and distribution exhibit natural monopoly characteristics.

Nonrival Energy: A term coined by Nic Carter to describe energy that is physically produced but cannot reach any willing consumer due to transmission constraints, geographic isolation, or timing mismatches between production and demand. Bitcoin mining can monetize nonrival energy by co-locating at the point of generation.

NERC CIP Standards: A set of mandatory cybersecurity standards developed by the North American Electric Reliability Corporation for the protection of critical infrastructure in the Bulk Electric

System. CIP standards address asset identification, electronic and physical security perimeters, personnel training, incident response, and supply chain risk management.

Net Metering: A billing arrangement in which customers with on-site generation (typically rooftop solar) receive credit on their electricity bills for excess energy exported to the grid, usually at the full retail rate.

North American Electric Reliability Corporation (NERC): The organization certified by FERC as the Electric Reliability Organization responsible for developing and enforcing mandatory reliability standards for the Bulk Electric System in North America.

Operating Reserve Demand Curve (ORDC): A scarcity pricing mechanism used in ERCOT that adds a price premium to real-time energy prices as a function of the probability of operating reserves falling below minimum required levels. Designed to provide investment incentives in an energy-only market.

Peaking Generation: Generation resources designed to operate only during periods of highest demand, characterized by low capital costs, high variable (fuel) costs, and the ability to start quickly. Typically natural gas combustion turbines (simple cycle gas turbines).

Power Usage Effectiveness (PUE): The ratio of total data center facility power to IT equipment power, used as a measure of data center energy efficiency. A PUE of 1.0 would mean all power goes to computing; industry average is approximately 1.5–1.6, with hyperscale operators achieving 1.1–1.2.

Public Utility Commission (PUC): A state-level regulatory body responsible for overseeing the rates, service quality, and operations of public utilities, including electric utilities, within the state.

Rate Base: The total value of a regulated utility's assets (net of depreciation) on which the utility is authorized to earn a rate of return. The foundation of cost-of-service regulation.

Regional Transmission Organization (RTO): An independent, FERC-regulated organization that coordinates, controls, and monitors the operation of the electric power system across a multi-state region, operates wholesale electricity markets, and ensures non-discriminatory access to the transmission grid.

Reliability Pricing Model (RPM): PJM's capacity market mechanism, which conducts forward auctions to procure capacity commitments for delivery three years in the future, using a downward-sloping demand curve to determine clearing prices and quantities.

Renewable Portfolio Standard (RPS): A state-level policy requirement that a specified percentage of electricity sold by utilities be generated from qualifying renewable energy sources. More than thirty states have adopted RPS requirements.

Reserve Margin: The percentage by which a system's available generation capacity exceeds its projected peak demand. A standard measure of resource adequacy; typical target reserve margins range from 13% to 17%.

Resource Adequacy: The ability of the electric power system to supply the aggregate electrical demand and energy requirements of end-use customers at all times, accounting for scheduled and unscheduled outages of system elements.

Retail Choice: A market structure in which end-use electricity customers can choose their electricity supplier from competing retail providers, rather than being required to purchase from their

local regulated utility. Also known as retail access, customer choice, or retail competition.

Security-Constrained Economic Dispatch (SCED): The optimization algorithm used by RTOs to determine the least-cost combination of generating resources to serve load while respecting all transmission constraints and reliability requirements. SCED produces Locational Marginal Prices as a byproduct.

Stranded Costs: Utility investments made under the traditional regulatory compact (e.g., expensive nuclear plants, above-market purchased power contracts) that may become uneconomic in a competitive market. The recovery of stranded costs was a major issue during electricity restructuring.

Swing Equation: The fundamental differential equation governing the rotational dynamics of a synchronous machine, relating the rate of change of rotor speed (and hence grid frequency) to the difference between mechanical power input and electrical power output.

Synchronous Generator: An AC generator in which the rotor's magnetic field rotates in precise synchrony with the electrical frequency of the grid. The traditional technology for converting mechanical energy to electrical energy at power plants, providing rotational inertia and other essential grid services.

Transmission System: The portion of the electric power system that transports electricity at high voltage (typically 69 kV and above) over long distances from generators to distribution substations. Includes transmission lines, substations, transformers, and associated equipment.

Value of Lost Load (VOLL): An administrative estimate of the economic cost to consumers of involuntary load shedding (blackouts), used in the design of scarcity pricing mechanisms and resource adequacy standards. Typically set in the range of $5,000 to $35,000 per megawatt-hour.

Vertically Integrated Utility (VIU): An electric utility that owns and operates generation, transmission, and distribution facilities, providing bundled service to end-use customers within a defined service territory under cost-of-service regulation.

Virtual Power Plant (VPP): A software platform and contractual structure that aggregates and coordinates the dispatch of many distributed energy resources (rooftop solar, home batteries, flexible loads) to function as a single dispatchable resource in wholesale markets.

Western Energy Imbalance Market (WEIM): A real-time energy market operated by CAISO that allows utilities across the Western Interconnection to buy and sell energy to balance supply and demand, without requiring full RTO membership.

* * *

Appendix C: Bibliography

Federal Regulatory Orders

- Federal Energy Regulatory Commission. *Order No. 888: Promoting Wholesale Competition Through Open Access Non-Discriminatory Transmission Services by Public Utilities.* Docket Nos. RM95-8-000 and RM94-7-001. Washington, DC: FERC, 1996.
- Federal Energy Regulatory Commission. *Order No. 2000: Regional Transmission Organizations.* Docket No. RM99-2-000. Washington, DC: FERC, 1999.
- Federal Energy Regulatory Commission. *Order No. 2222: Participation of Distributed Energy Resource Aggregations in Markets Operated by Regional Transmission Organizations and Independent System Operators.* Docket No. RM18-9-000. Washington, DC: FERC, September 17, 2020.
- Federal Energy Regulatory Commission. *Order No. 1920: Building for the Future Through Electric Regional Transmission Planning and Cost Allocation.* Docket No. RM21-17-000. Washington, DC: FERC, 2024.
- Federal Energy Regulatory Commission. *Order No. 2023: Improvements to Generator Interconnection Procedures and Agreements.* Docket No. RM22-14-000. Washington, DC: FERC, 2023.

Federal Statutes

- Federal Power Act, 16 U.S.C. §§ 791a–828c (1935).
- Tennessee Valley Authority Act, 16 U.S.C. § 831 et seq. (1933).
- Rural Electrification Act of 1936, 7 U.S.C. § 901 et seq.
- Bonneville Project Act, 16 U.S.C. § 832 et seq. (1937).

- Public Utility Holding Company Act of 1935, 15 U.S.C. §§ 79a–79z-6.
- Public Utility Regulatory Policies Act of 1978 (PURPA), 16 U.S.C. § 2601 et seq.
- Energy Policy Act of 2005, Pub. L. 109-58, 119 Stat. 594.
- Infrastructure Investment and Jobs Act, Pub. L. 117-58, 135 Stat. 429 (2021).
- Inflation Reduction Act of 2022, Pub. L. 117-169, 136 Stat. 1818.

Reliability Standards

- North American Electric Reliability Corporation. *CIP-002 through CIP-013: Critical Infrastructure Protection Standards.* Atlanta, GA: NERC.
- North American Electric Reliability Corporation. *TPL-007-4: Transmission System Planned Performance for Geomagnetic Disturbance Events.* Atlanta, GA: NERC.
- Institute of Electrical and Electronics Engineers. *IEEE Standard 1366-2022: Guide for Electric Power Distribution Reliability Indices.* Piscataway, NJ: IEEE, 2022. Defines SAIDI, SAIFI, and related distribution reliability metrics. (Figure 1.2.)
- Institute of Electrical and Electronics Engineers. *IEEE Standard 2800-2022: Standard for Interconnection and Interoperability of Inverter-Based Resources Interconnecting with Associated Transmission Electric Power Systems.* Piscataway, NJ: IEEE, 2022.

Academic Papers and Research Reports

- Averch, Harvey, and Leland L. Johnson. "Behavior of the Firm Under Regulatory Constraint." *American Economic Review* 52, no. 5 (December 1962): 1052–1069.
- Anderson, P. M., and A. A. Fouad. *Power System Control and Stability.* 2nd ed. Piscataway, NJ: IEEE Press / Wiley-Interscience, 2003. (Figure 10.3.)
- Carter, Nic, and Brad Jones. "Leveraging Bitcoin Miners as Flexible Load Resources for Power System Stability and Efficiency." Castle Island Ventures and ERCOT.
- Carter, Nic, and Ross Stevens. "Bitcoin Net Zero." NYDIG Research.
- Rhodes, Joshua D., and Thomas Deetjen. "Analysis of Flexible Data Centers on the ERCOT Grid." IdeaSmiths LLC.
- Alden, Lyn. "Bitcoin's Energy Usage Isn't a Problem. Here's Why." LynAlden.com.
- Arcane Research (K33). *Co-Located Mining and Wind Generation: Case Studies and Economic Analysis.* Oslo: Arcane Research, 2022. Foreword by Nic Carter.
- "Valorization of Curtailed Power: Enhancing Grid Flexibility Through Cryptocurrency Mining." *Journal of Climate Change Research.*

- Cambridge Centre for Alternative Finance. *Cambridge Bitcoin Electricity Consumption Index.* Cambridge: University of Cambridge, ongoing.
- Lazard. *Lazard's Levelized Cost of Energy Analysis.* New York: Lazard. Published annually. (Figure 6.2.)
- Federal Energy Regulatory Commission. *Energy Primer: A Handbook of Energy Market Basics.* Washington, DC: FERC, 2020. (Figure 6.2.)

Government Reports and Studies

- Congressional EMP Commission. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack.* Washington, DC, 2004; reconstituted 2015.
- U.S. Department of Energy. *National Transmission Needs Study.* Washington, DC: DOE.
- National Renewable Energy Laboratory. *Transmission Expansion Studies for Renewable Integration.* Golden, CO: NREL.
- U.S. Energy Information Administration. *Cryptocurrency Mining Energy Consumption Data Collection.* Washington, DC: EIA.
- Lawrence Berkeley National Laboratory. *United States Data Center Energy Usage Report.* Berkeley, CA: LBNL. (Figure 12.1.)
- Goldman Sachs Research. *AI, Data Centers and the Coming U.S. Power Demand Surge.* Goldman Sachs Global Investment Research, 2024. (Figure 12.1.)

Statistical Data Sources

- U.S. Energy Information Administration. *Form EIA-861: Annual Electric Power Industry Report.* Washington, DC: EIA. Published annually. (Figures 1.0, 1.2.)
- U.S. Energy Information Administration. *Form EIA-923: Power Plant Operations Report.* Washington, DC: EIA. Published monthly and annually. (Figures 1.1, 10.1.)
- U.S. Energy Information Administration. *Electric Power Annual.* Washington, DC: EIA. Published annually. (Figures 3.1, 3.2, 6.1, 8.1.)
- U.S. Energy Information Administration. *Short-Term Energy Outlook (STEO).* Washington, DC: EIA. Published monthly. (Figures 1.0, 10.1.)
- U.S. Energy Information Administration. *Annual Energy Outlook.* Washington, DC: EIA. Published annually.
- North American Electric Reliability Corporation. *Long-Term Reliability Assessment.* Atlanta, GA: NERC. Published annually. (Figure 6.1.)
- California Independent System Operator. *Managing Oversupply* (original Duck Curve

projections). Folsom, CA: CAISO, 2013. (Figure 8.2.)

- California Independent System Operator. *Annual Report on Market Issues and Performance.* Folsom, CA: CAISO. Published annually. (Figures 8.2, 10.3.)
- Edison Electric Institute. *Statistical Yearbook of the Electric Power Industry.* Washington, DC: EEI. Published annually. (Figure 10.2.)
- Federal Energy Regulatory Commission. *Form 1: Electric Utility Annual Report.* Washington, DC: FERC. Published annually. (Figure 10.2.)
- Electric Reliability Council of Texas. *Fact Sheet and Quick Facts.* Austin, TX: ERCOT. Published periodically. (Figure 13.1.)

Industry Publications

- Electricity Information Sharing and Analysis Center (E-ISAC). *Threat Intelligence Reports.* Washington, DC: E-ISAC.
- Uptime Institute. *Tier Classification System for Data Center Infrastructure.* New York: Uptime Institute.
- World Bank. *Global Gas Flaring Reduction Partnership: Annual Reports.* Washington, DC: World Bank Group.

Cybersecurity Incident References

- SANS Industrial Control Systems. "Analysis of the Ukraine Power Grid Attack: December 2015." SANS ICS, 2016.
- Dragos, Inc. "CRASHOVERRIDE: Analysis of the Threat to Electric Grid Operations." Dragos Security, 2017.
- FireEye (Mandiant). "Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims." FireEye Threat Research, December 2020.
- Cybersecurity and Infrastructure Security Agency (CISA). "Volt Typhoon Targets U.S. Critical Infrastructure." Advisory AA24-038A. Washington, DC: CISA, 2024.

Historical Events

- U.S.-Canada Power System Outage Task Force. *Final Report on the August 14, 2003 Blackout in the United States and Canada.* Washington, DC and Ottawa, ON, 2004.

- Federal Energy Regulatory Commission and North American Electric Reliability Corporation. *The February 2021 Cold Weather Outages in Texas and the South Central United States.* FERC-NERC-Regional Entity Staff Report. Washington, DC: FERC, November 2021.
- National Hurricane Center. *Tropical Cyclone Reports: Hurricane Maria (2017), Hurricane Harvey (2017), Hurricane Ian (2022).* Miami, FL: NHC/NOAA.
- California Public Utilities Commission. *Investigation of Utility Wildfire Practices: Pacific Gas and Electric Company.* San Francisco, CA: CPUC, 2019.

* * *

Index

V

W

About the Author

PRESTON P. PRATT currently serves as an Area Director of Facility Operations for a leading hyperscale technology company, overseeing data center infrastructure across a major region of North America. In that role, he is responsible for the physical infrastructure and operational continuity of some of the most power-intensive facilities on the planet — the hyperscale data centers and AI systems at the leading edge of a global infrastructure buildout that is transforming how the world computes, communicates, and consumes electricity.

Pratt spent more than a decade at the Tennessee Valley Authority, the largest public power provider in the United States. As Plant Manager of Sequoyah Nuclear Plant — a 2,400-megawatt, two-unit Westinghouse pressurized water reactor station — he led approximately 800 employees responsible for the safe and reliable operation of one of the largest generating assets in the TVA fleet. He later served as Director of Strategic Operational Solutions at TVA's corporate office, where he led initiatives spanning digital transformation, carbon reduction planning, and generation capacity strategy, giving him a perspective on the grid not only from the plant floor but from the level where resource and investment decisions are made.

During a loaned assignment to the Institute of Nuclear Power Operations, Pratt served as a Performance Recovery Leader and Performance Monitoring Leader, working directly with nuclear stations across the United States on organizational effectiveness, leadership, and operational fundamentals. The assignment gave him firsthand exposure to a broad cross-section of the American nuclear fleet — its operational diversity, its common challenges, and the institutional culture that sustains it.

Earlier in his career, Pratt worked at Pennsylvania Power & Light's Susquehanna Steam Electric Station within the PJM Interconnection, where he served as a Senior Reactor Operator. He has been licensed by the Nuclear Regulatory Commission three times — on both boiling water and pressurized water reactor designs as well as a research reactor — a credential that reflects the depth of his technical foundation and the standards of accountability the nuclear industry demands.

Pratt's career began in the United States Navy's nuclear propulsion program, where he served aboard USS George Washington and earned the Navy and Marine Corps Achievement Medal. He holds a Master of Science in Nuclear Engineering and a Bachelor of Science in Mechanical Engineering, both from Texas A&M University. He is a fellow of both the Institute of Nuclear Power Operations and the Department of Energy's Advanced Fuel Cycle Initiative, a graduate of Vanderbilt University's leadership program, and served an elected term on the Executive Committee of the American Nuclear Society's Operations and Power Division.

The arc from Navy reactor compartment to nuclear control room to corporate strategy to hyperscale data center is not a coincidence. Pratt recognized early that the explosive growth of artificial intelligence and cloud computing would create the most significant shift in electricity demand since the postwar industrial boom — and that the people best equipped to build and operate that infrastructure were those who already understood power systems at their most demanding. In Proof of Power, he brings both halves of the grid's modern challenge into focus: the utility infrastructure built to power a different era, and the digital economy now demanding far more of it than anyone planned for.

Connect with the Author

prestonppratt.com

Blog: blog.prestonppratt.com
LinkedIn: linkedin.com/in/prestonppratt
X / Twitter: x.com/prestonppratt
Facebook: facebook.com/prestonppratt

Published by Pratt Materials • prattmaterials.com

For bulk orders, speaking inquiries, and media requests,
please visit prestonppratt.com.

PROOF OF POWER

The American electric grid is the largest machine ever built—a continent-spanning network of generators, transmission lines, and distribution systems that must balance supply and demand every fraction of a second. It is simultaneously a marvel of engineering, a product of a century of political bargaining, and an institution under unprecedented stress.

Proof of Power is the first comprehensive account of this system in all its dimensions: the physics that constrain it, the markets that price it, the regulations that govern it, and the technologies that are transforming it. From the synchronous generators that maintain 60 Hz frequency to the inverter-based resources that threaten to destabilize it, from the capacity markets of PJM to the energy-only experiment of ERCOT, from NERC CIP cybersecurity standards to the explosive demand of hyperscale data centers and Bitcoin mining facilities—this book makes the grid legible.

Written for policymakers, engineers, investors, and anyone who has ever wondered what happens between the power plant and the light switch, Proof of Power is the definitive guide to the infrastructure that powers modern civilization.

"From Navy reactor compartment to nuclear control room
to corporate strategy to hyperscale data center."
— prestonppratt.com

ISBN 000-0-00-000000-0
prestonppratt.com