

Appendix B: Electric Power System Basics

B.1 INTRODUCTION

Electricity is critical to our daily lives, yet most people have little understanding of the complex process that brings electric power to our homes, offices, and factories whenever we demand it. This appendix is a tutorial on how the electric power system works. We assume no prior knowledge in the area and start by providing a description of the physical foundations of electricity. We next discuss the structure and components of the electric power system. We follow with an explanation of how the system is operated and how wholesale electricity markets work. In the final section, we provide a brief overview of system planning. Because there are slight differences in the structure, operation, and planning of the electric power system from country to country and region to region, we focus mostly on fundamental aspects that remain unchanged; however, where appropriate we provide U.S.-centric details and highlight important variations in practice.

B.2 FUNDAMENTALS OF ELECTRIC POWER

To understand electric power systems, it is helpful to have a basic understanding of the fundamentals of electricity. These include the concepts of energy, voltage, current, direct current (dc), alternating current (ac), impedance, and power.ⁱ

Energy

Energy is the ability to perform work. Energy cannot be created or destroyed but can be converted from one form to another.ⁱⁱ For example, chemical energy in fossil fuels can be converted into electrical energy, and electrical energy in turn can be converted into useful work in the form of heat, light, and motion. While the scientific community measures energy in watt-seconds or joules, traditionally in the electric power industry, energy is measured in watt-hours (Wh) and for larger values is expressed in kilowatt (thousand watt, kW), megawatt (million watt, MW), gigawatt (billion watt, GW), or terawatt (trillion watt, TW) hours.ⁱⁱⁱ A 100 watt lightbulb consumes 2,400 Wh (or 2.4 kWh) of energy in 24 hours, and the total annual electrical energy consumption of the U.S. in 2010 was about 3,900 TWh.¹ One kilowatt hour is equivalent to 3.6 megajoules.

Voltage

Voltage (also referred to as potential) is measured between two points and is a measure of the capacity of a device connected to those points to perform work per unit of charge that flows between those points. Voltage can be considered analogous to the pressure in a water pipe. Voltage is measured in volts (V), and for large values expressed in kilovolts (kV) or megavolts (MV).

ⁱ Those who only desire a high-level understanding of electric power systems can skip this section.

ⁱⁱ If mass is not considered a form of energy, an exception is in nuclear reactions, where mass and energy can be transformed into one another.

ⁱⁱⁱ Watt is the unit of power, or the rate of flow (or consumption) of energy, as discussed later in this section.

Current

Current is a measure of the rate of flow of charge through a conductor. It is measured in amperes. Current can be considered analogous to the rate of flow of water through a pipe.

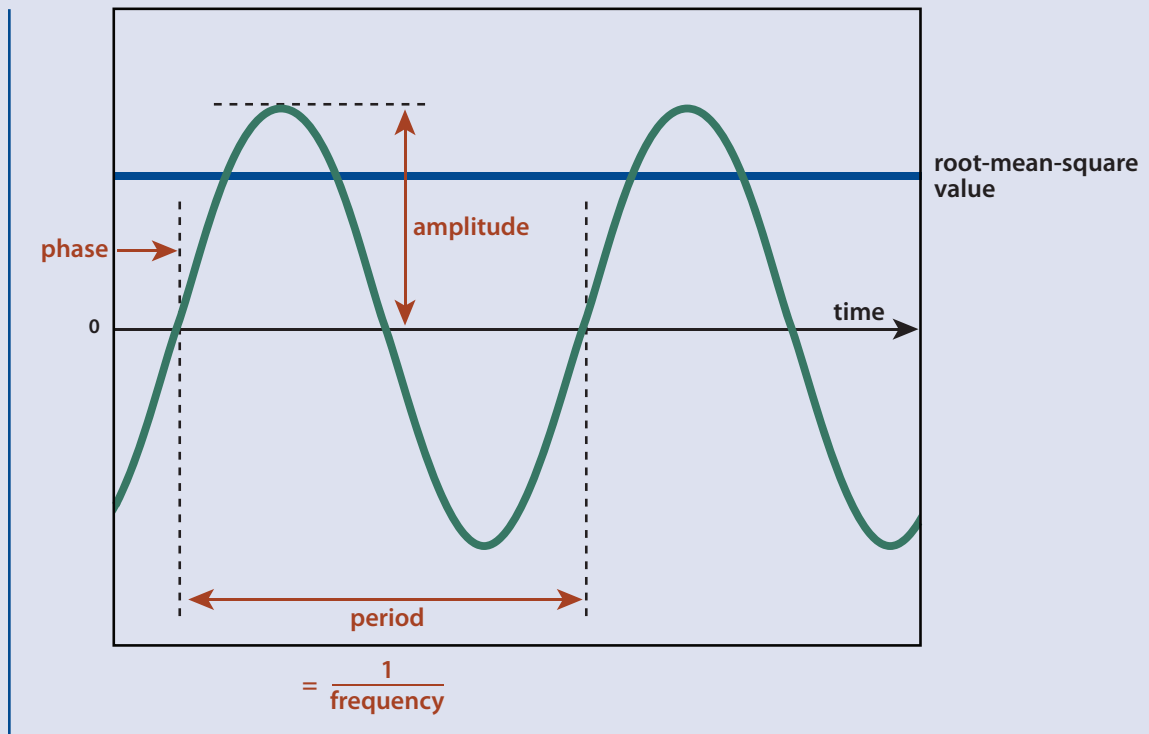
Dc and Ac

Current can be unidirectional, referred to as “direct current,” or it can periodically reverse directions with time, in which case it is called “alternating current.” Voltage also can be unipolar—in which one point is always at a higher voltage than the other—or alternating in polarity with time. Unipolar voltage is referred to as “dc voltage.” Voltage that reverses polarity in a periodic fashion is referred to as “ac voltage.” Alternating currents and voltages in power systems have nearly sinusoidal profiles.

Ac voltage and current waveforms are defined by three parameters: amplitude, frequency, and phase, as shown in Figure B.1. The maximum value of the waveform is referred to as its “amplitude.” The amplitude of the ac voltage in a standard 120 V outlet is 170 V. The 120 V in this case refers to the root-mean-square (rms) value of the voltage and is the equivalent dc voltage with the capacity to perform the same amount of work. In the case of ac, the amplitude is equal to the rms value multiplied by the square root of two. In the case of dc, the amplitude and rms values are the same.

Frequency is the rate at which current and voltage in the system oscillate, or reverse direction and return. Frequency is measured in cycles per second, also called “hertz” (Hz). In the U.S., as well as the rest of North America and parts of South America and Japan, the ac system frequency is 60 Hz, while in the rest of the world it is 50 Hz.² Dc can be considered a special case of ac, one with frequency equal to zero.

Figure B.1 Amplitude, Frequency, Period, and Phase of an Alternating Current or Voltage Waveform



The time in seconds it takes for an ac waveform to complete one cycle, the inverse of frequency, is called the “period.” The phase of an ac waveform is a measure of when the waveform crosses zero relative to some established time reference. Phase is expressed as a fraction of the ac cycle and measured in degrees (ranging from -180 to +180 degrees). There is no concept of phase in a dc system.

Electric power systems are predominantly ac, although a few select sections are dc. Ac is preferred because it allows voltage levels to be changed with ease using a transformer. The voltage level of a dc system also can be changed, but doing so requires more sophisticated and expensive equipment using power electronics technology. However, dc can be advantageous when energy has to be transmitted over long distances for reasons discussed later. Dc also is used to connect ac systems that operate at different frequencies (as in Japan) or systems with identical frequencies that are not synchronized (as between interconnections in the U.S.).^{iv}

Impedance

Impedance is a property of a conducting device—for example, a transmission line—that represents the impediment it poses to the flow of current through it. The rate at which energy flows through a transmission line is limited by the line’s impedance. Impedance has two components: resistance and reactance. Impedance, resistance, and reactance are all measured in ohms.

Resistance

Resistance is the property of a conducting device to resist the flow of ac or dc current through it. A transmission line is composed of wires known as “conductors” whose resistance increases with length and decreases with increasing conductor cross-sectional area.

Resistance causes energy loss in the conductor as moving charges collide with the conductor’s atoms and results in electrical energy being converted into heat. However, resistance does not introduce any phase shift between voltage and current. The rate of energy loss (called “power loss”) is equal to the resistance times the square of the rms current.

Reactance

Voltages and currents create electric and magnetic fields, respectively, in which energy is stored. Reactance is a measure of the impediment to the flow of power caused by the creation of these fields. When the voltage and current are ac, this alternating storage and retrieval of energy retards the flow of power but no energy is lost. When energy is stored in magnetic fields, the element is said to have “inductive reactance,” while “capacitive reactance” describes elements creating energy stored in electric fields. Reactance is a function of frequency—inductive reactance increases with frequency while capacitive reactance decreases. The presence of reactance in a system also creates a phase shift between voltage and current—inductive reactance causes the current to lag the voltage (a negative phase shift), while capacitive reactance forces the current to lead the voltage (a positive phase shift). (One way to visualize this is that the current is “busy” storing energy in a magnetic field as the voltage proceeds, while the voltage is “busy” storing it in an electric field as the current proceeds.)

The impedance of a transmission line is primarily comprised of inductive reactance. Therefore its current will be out of phase with and lag its voltage, which is undesirable for reasons discussed later. To compensate for this, elements with capacitive reactance (capacitors) are connected to the transmission line. The positive phase shift caused by these capacitors cancels out the negative shift due to the inductive

^{iv} Synchronized systems are at the same frequency and have a specific phase difference between their voltages.

reactance of the transmission line and forces the transmission voltage and current to be in phase (a good thing). This process is called “line compensation.”

The inductive reactance of a transmission line is proportional to both frequency and line length, and for long ac lines the inductive reactance limits the amount of power the line can carry. At zero frequency (dc) the reactance is zero, making dc attractive for long-distance transmission.

Power

Power is the rate at which energy is flowing or work is being done.^v Since voltage is the amount of work done for each unit of charge that flows and current is the rate of flow of charge, the product of voltage and current is the rate of work—power, or more precisely instantaneous power. Since power loss is equal to the resistance of a conductor times the square of the current, loss in a transmission line can be reduced by increasing the transmission voltage, which allows the current to be reduced for the same amount of power transmitted. As a result, long transmission lines employ high voltage. However, as discussed later, high-voltage lines also have drawbacks, including the need to maintain larger clearances to maintain safety.

In ac systems, where voltage and current oscillate many times a second, the instantaneous power they produce is also rapidly varying, as shown in Figure B.2. In the figure, negative instantaneous power is equivalent to power flowing in the backwards direction. In electric power systems, it is more valuable to have measures of power that are averages over many cycles. These measures are real power, reactive power, and apparent power. Only two of these three measures are independent; apparent power can be determined from real power and reactive power.

Real Power

Real power, also called “active power” or “average power,” is the average value of instantaneous power, as shown in Figure B.2, and is power that actually does work. It is measured in watts. Although instantaneous power can be flowing in both directions, real power only flows in one direction, as shown in Figure B.2(a)–(c). Real power is zero if the phase difference between voltage and current is 90 degrees, as shown in Figure B.2(d).

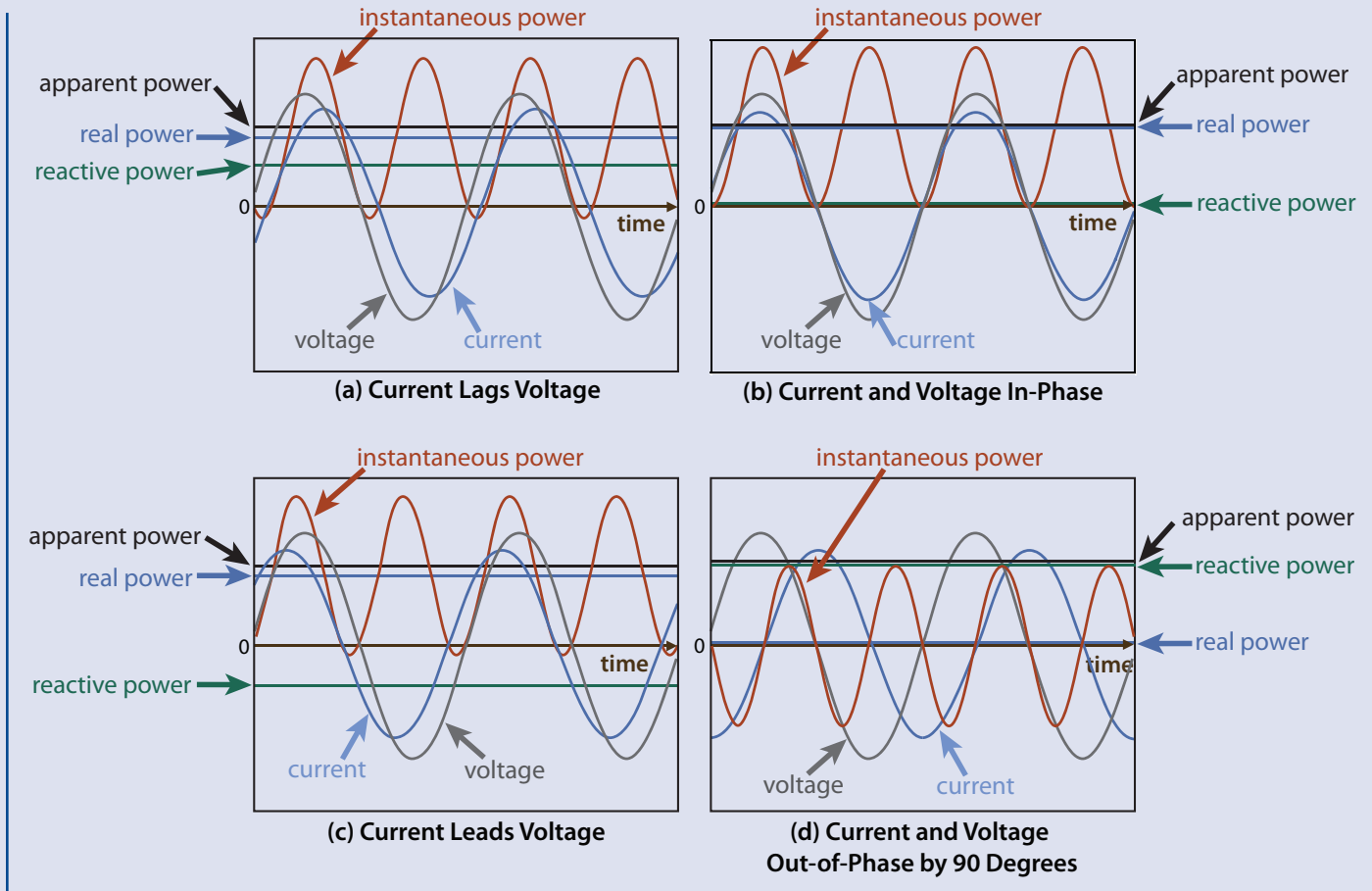
Reactive Power

If the voltage and current waveforms are “in phase”—that is, they cross zero at the same time—then instantaneous power, although varying, is always positive or flowing in one direction (Figure B.2(b)). In this case, all the power is real power. However, if one waveform is shifted in time relative to the other, a condition called “out of phase,” then power takes on both positive and negative values, as shown in Figure B.2(a), (c), and (d). This phase difference can arise, for example, because of the reactance of the transmission line. Here, in addition to the real power that is flowing in one direction, there is back and forth movement of power called “reactive power.” While it does no useful work, reactive power flow still causes power losses in the system because current is flowing through components, such as transformers and transmission lines, which have resistance. Reactive power is measured in volt-amperes reactive (VAR).

Reactive power can be positive or negative. But unlike instantaneous power, its sign does not indicate the direction of reactive power flow. Instead, the sign simply indicates the relative phase shift between current and voltage. When current lags voltage due to the presence of inductive reactance, reactive power is positive, as shown in Figure B.2(a); when current leads voltage due to the presence of capacitive

^v It is energy that actually “flows” in a power system, power being the rate of this energy flow. However, though technically incorrect, common usage is to speak of “power flow.”

Figure B.2 Current, Voltage, and Power in an Ac System



reactance, reactive power is negative, as shown in Figure B.2(c). Equipment that draws negative reactive power is often said to be “supplying” reactive power. In power systems, capacitors are often connected near large inductive loads to compensate for their positive reactive power.

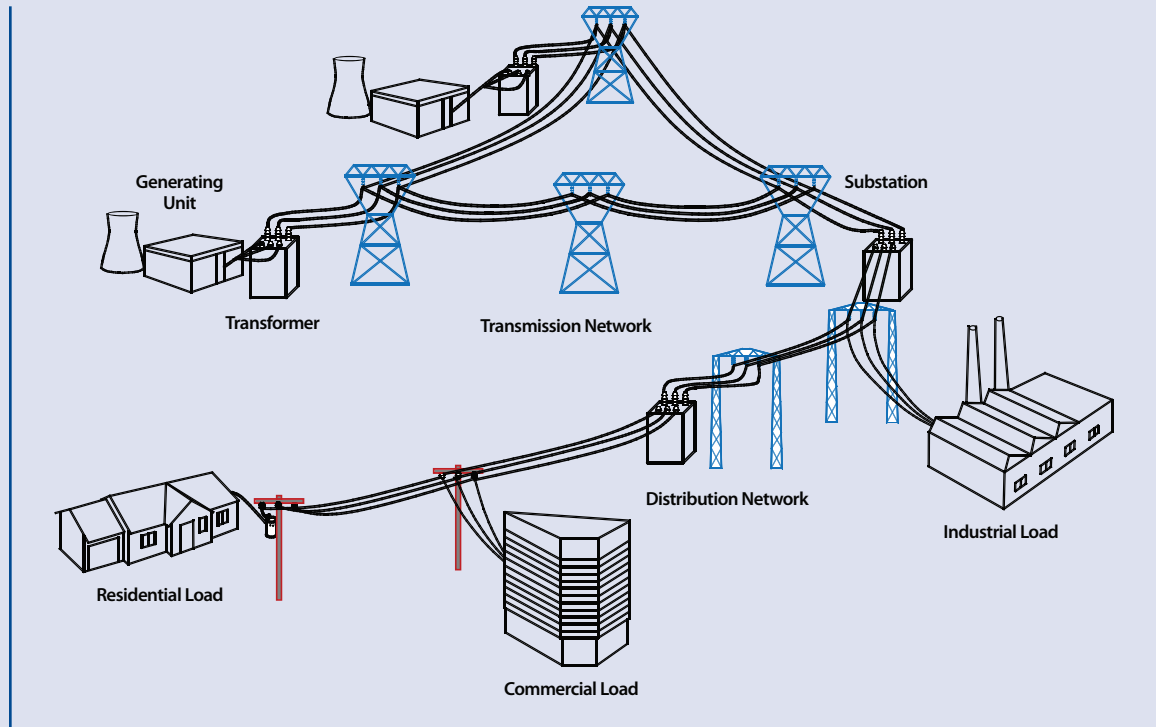
Apparent Power

Apparent power is the product of rms voltage and rms current, and is always greater than or equal to real and reactive power. Electrical equipment, such as transformers and transmission lines, must be thermally rated for the apparent power they process. Apparent power is measured in volt amperes. The ratio of real power to apparent power is called “power factor.” Utilities like to maintain a unity power factor as it implies that all of the power that is flowing is doing useful work.

B.3 STRUCTURE OF THE ELECTRIC POWER SYSTEM

The electric power system consists of generating units where primary energy is converted into electric power, transmission and distribution networks that transport this power, and consumers’ equipment (also called “loads”) where power is used. While originally generation, transport, and consumption of electric power were local to relatively small geographic regions, today these regional systems are connected together by high-voltage transmission lines to form highly interconnected and complex systems that span wide areas. This interconnection allows economies of scale, better utilization of the most economical

Figure B.3 Structure of the Electric Power System



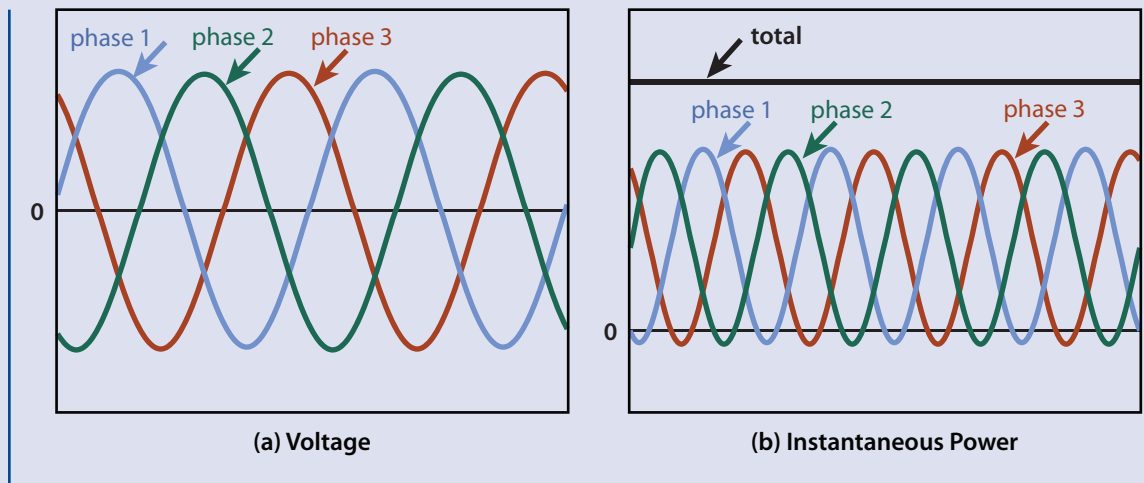
generators, increased reliability, and an improved ratio of average load to peak load due to load diversity, thus increasing capacity utilization. Interconnection also leads to complexity, however, as any disturbance in one part of the system can adversely impact the entire system. Figure B.3 illustrates the basic structure of the electric power system. We discuss each of its subsystems next.

Generation

Electric power is produced by generating units, housed in power plants, which convert primary energy into electric energy. Primary energy comes from a number of sources, such as fossil fuel and nuclear, hydro, wind, and solar power. The process used to convert this energy into electric energy depends on the design of the generating unit, which is partly dictated by the source of primary energy.

The term “thermal generation” commonly refers to generating units that burn fuel to convert chemical energy into thermal energy, which is then used to produce high-pressure steam. This steam then flows and drives the mechanical shaft of an ac electric generator that produces alternating voltage and current, or electric power, at its terminals. These generators have three terminals and produce three ac voltages, one at each terminal, which are 120 degrees out of phase with respect to each other, as shown in Figure B.4(a). This set of voltages is known as “three-phase ac voltage,” whereas the voltage discussed in the previous section and illustrated in Figure B.1 is known as “single-phase ac voltage.” Three-phase ac has multiple advantages over single-phase ac, including requiring less conducting material in the transmission lines and allowing the total instantaneous power flowing from the generator to be constant (Figure B.4(b)).

Figure B.4 Three-Phase System



Nuclear generating units use an energy conversion process similar to thermal units, except the thermal energy needed to produce steam comes from nuclear reactions. Hydro and wind generating units convert the kinetic energy of water and wind, respectively, directly into rotation of the electric generator's mechanical shaft. Solar-thermal and geothermal generating units use the sun's radiation and the Earth's warmth, respectively, to heat a fluid and then follow a conversion process similar to thermal units. Solar photovoltaic generating units are quite different and convert the energy in solar radiation directly into electrical energy. Another common type of generating unit is the gas, or combustion, turbine. These burn a pressurized mixture of natural gas and air in a jet engine that drives the electric generator. Combined-cycle gas turbine plants have a gas turbine and a steam turbine. They reuse the waste heat from the gas turbine to generate steam for the steam turbine and hence achieve higher energy conversion efficiencies.^{vi}

From the operational perspective of the electric power system, generating units are classified into three categories: baseload, intermediate,

and peaking units. Baseload units are used to meet the constant, or base, power needs of the system. They run continuously throughout the year except when they have to be shut down for repair and maintenance. Therefore, they must be reliable and economical to operate. Because of their low fuel costs, nuclear and coal plants are generally used as baseload units, as are run-of-the-river hydroelectric plants. However, nuclear and coal baseload units are expensive to build and have slow ramp rates—that is, their output power can be changed only slowly (on the order of hours).

Intermediate units, also called cycling units, operate for extended periods of time but, unlike baseload units, not at one power continuously. They have the ability to vary their output more quickly than baseload units. Combined-cycle gas turbine plants and older thermal generating units generally are used as intermediate units.

Peaking units operate only when the system power demand is close to its peak. They have to be able to start and stop quickly, but they run only for a small number of hours in a year. Gas

^{vi} Combined-cycle plants can have efficiencies in the 55%–60% range, compared to about 40% for conventional thermal plants.

turbine and hydroelectric plants with reservoirs are generally used as peaking units. Gas turbines are the least expensive to build but have high operating costs.

Large generating units generally are located outside densely populated areas, and the power they produce has to be transported to load centers. They produce three-phase ac voltage at the level of a few to a few tens of kV. To reduce power losses during onward transmission, this voltage is immediately converted to a few hundred kV using a transformer. All the generators on a single ac system are synchronized.

In addition to the main large generating units, the system typically also has some distributed generation, including combined heat and power units. These and other small generating units, such as small hydroelectric plants, generally operate at lower voltages and are connected at the distribution system level. Small generating units, such as solar photovoltaic arrays, may be single-phase.

Transmission

The transmission system carries electric power over long distances from the generating units to the distribution system. The transmission network is composed of power lines and stations/substations. Transmission system power lines, with rare exceptions, are attached to high towers. However, in cities, where real estate is valuable, transmission lines are sometimes made up of insulated cables buried underground. Stations and substations house transformers, switchgear, measurement instrumentation, and communication equipment. Transformers are used to change the level of the transmission voltage. Switchgear includes circuit breakers and other types of switches used to disconnect parts of the transmission network for system protection or maintenance.

Measurement instrumentation collects voltage, current, and power data for monitoring, control, and metering purposes. Communication equipment transmits these data to control centers and also allows switchgear to be controlled remotely.

Since transmission networks carry power over long distances, the voltage at which they transmit power is high to reduce transmission losses, limit conductor cross-sectional area, and require narrower rights-of-way for a given power. However, to maintain safety, high transmission voltages require good insulation and large clearance from the ground, trees, and any structures. Transmission voltages vary from region to region and country to country. The transmission voltages commonly (but not exclusively) used in the U.S. are 138 kV, 230 kV, 345 kV, 500 kV, and 765 kV.³ A voltage of 1,000 kV has been used on a transmission line in China. Although most transmission is three-phase ac, for very-long-distance transmission, HVDC can be beneficial because transmission lines present no reactive impedance to dc. HVDC also only requires two conductors instead of three. However, HVDC transmission lines require expensive converter stations (utilizing power electronics technology) at either end of the line to connect to the rest of the ac system.

Transformers at transmission substations convert transmission voltages down to lower levels to connect to the subtransmission network or directly to the distribution network. Subtransmission carries power over shorter distances than transmission and is typically used to connect the transmission network to multiple nearby relatively small distribution networks. In the U.S., the commonly used subtransmission voltages are 69 kV and 115 kV.

Topologically, the transmission and subtransmission line configurations are mesh networks (as opposed to radial), meaning there are multiple paths between any two points on the network. This redundancy allows the system to provide power to the loads even when a transmission line or a generating unit goes offline. Because of these multiple routes, however, the power flow path cannot be specified at will. Instead power flows along all paths from the generating unit to the load. The power flow through a particular transmission line depends on the line's impedance and the amplitude and phase of the voltages at its ends,^{vii} as discussed in Box B.1. Predicting these flows requires substantial computing power and precise knowledge of network voltages and impedances, which are rarely known with high precision. Hence, precise prediction of the power flowing down a particular transmission line is difficult. The presence of multiple paths between generation and load in the transmission network also leads to flows on undesirable paths. These undesirable flows are known as “loop flows.”

The power that can be transmitted on a transmission line is limited by either thermal, voltage stability, or transient stability constraints, depending on which is the most binding, as illustrated in Figure 2.4 in Chapter 2.^{viii, 4} The thermal constraint arises due to the resistance of the transmission line that causes excessive power losses and hence heating of the line when the power flowing through it exceeds a certain level. The voltage stability constraint arises due to the reactance of a transmission line that causes the voltage at the far end of the line to drop below an allowable level (typically 95% of the nominal design voltage level) when the power flowing through the line exceeds a certain level. The transient stability constraint relates to the ability of the transmission line to

deal with rapid changes in the power flowing through it without causing the generators to fall out of synchronism with each other. Generally, maximum power flow on short transmission lines is limited by thermal constraints, while power flow on longer transmission lines is limited by either voltage or transient stability constraints. These power flow constraints cause so-called congestion on transmission lines, when the excess capacity in the lowest-cost generating units cannot be supplied to loads due to the limited capacity of one or more transmission lines.

Some very large consumers take electric power directly from the transmission or subtransmission network. However, the majority of consumers get their power from the distribution network.

Distribution

Distribution networks carry power the last few miles from transmission or subtransmission to consumers. Power is carried in distribution networks through wires either on poles or, in many urban areas, underground. Distribution networks are distinguished from transmission networks by their voltage level and topology. Lower voltages are used in distribution networks, as lower voltages require less clearance. Typically lines up to 35 kV are considered part of the distribution network.

The connection between distribution networks and transmission or subtransmission occurs at distribution substations. Distribution substations have transformers to step voltage down to the primary distribution level (typically in the 4 to 35 kV range in the U.S.). Like transmission substations, distribution substations also have circuit breakers and monitoring equipment. However, distribution substations are generally less automated than transmission substations.

^{vii} The power flow through a transmission line is roughly proportional to the phase difference between the voltages at its ends and inversely proportional to its impedance.

^{viii} Power system stability limits also are discussed in Box 2.3 in Chapter 2.

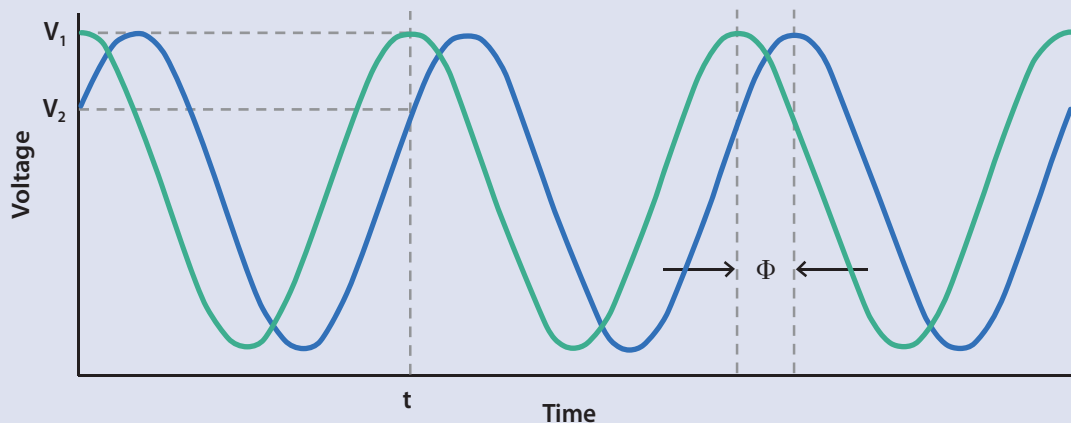
BOX B.1 CONTROLLING POWER FLOW

Two factors determine power flow: the impedance of a line and the difference in the instantaneous voltages at its two ends. Impedance is the combination of resistance and reactance. Resistance accounts for energy that is lost as heat in the line. It is analogous to the physical resistance exerted by water on a swimmer or wind on a cyclist. Energy lost in this way can never be recovered. Reactance accounts for energy associated with the electric and magnetic fields around the line. This energy is analogous to the potential energy stored when riding a bicycle up a hill. It is recovered (in the ideal case) when going down the other side. In an alternating current (ac) line in the U.S., this energy is stored and recovered 120 times per second, and thus is quite different from the behavior of energy stored in devices such as batteries. The resistance of a line is determined by the material properties, length, and cross-section of the conductor, while reactance is determined by geometric properties (the position of conductors relative to each other and ground). In practical transmission lines, resistance is small compared to reactance, and thus reactance has more influence on power flow than resistance.

As a function of time, the voltages at the ends of a transmission line are sinusoidal in shape. In the figure below, the two sinusoids represent voltages at opposite ends of a line. When there is power flow, the instantaneous values of voltage at the two ends of the line are different, as shown by the difference in voltages (V_1 and V_2) at time (t) in the figure. This instantaneous difference is a function of the difference in phase angle between the two sinusoids. The phase angle difference is shown in the figure as ϕ . If the two voltages are in phase, that is, if $\phi = 0$, then there will be no difference in their instantaneous values.

The power flow on a line varies directly with the phase angle difference (or more precisely the sine of the phase angle difference) and inversely with the line's impedance. Except in very special cases in which devices are used to control power flow on individual lines, the flow of power in a line is difficult to control when the line is part of an interconnected network since the characteristics of the entire network collectively determine power flows. When special devices are used to control power flow, they do so by modifying impedance and phase angle.

Phase Angle Difference (ϕ) of Voltage Sinusoids at the Ends of a Transmission Line



Primary distribution lines leaving distribution substations are called “feeders.” They also carry three-phase ac voltage, which is why one sees three wires on many poles in rural and suburban areas. These individual phases are then separated and feed different neighborhoods.

Distribution networks usually have a radial topology, referred to as a “star network,” with only one power flow path between the distribution substation and a particular load. Distribution networks sometimes have a ring (or loop) topology, with two power flow paths between the distribution substation and the load. However, these are still operated as star networks by keeping a circuit breaker open. In highly dense urban settings, distribution networks also may have a mesh network topology, which may be operated as an active mesh network or a star network. The presence of multiple power flow paths in ring and mesh distribution networks allows a load to be serviced through an alternate path by opening and closing appropriate circuit breakers when there is a problem in the original path. When this process is carried out automatically, it is often referred to as “self-healing.” Distribution networks usually are designed assuming power flow is in one direction. However, the addition of large amounts of distributed generation may make this assumption questionable and require changes in design practices.

Industrial and large commercial users usually get three-phase supply directly from the primary distribution feeder, as they have their own transformers and in certain cases can directly utilize the higher voltages. However, for the remaining consumers, who generally require only single-phase power, power is usually transmitted for the last half-mile or so over lateral feeders that carry one phase. A distribution transformer, typically mounted on a pole or located underground near the customer, steps this voltage down to the secondary distribution level, which is safe

enough for use by general consumers. Most residential power consumption in the U.S. occurs at 120 V or 240 V. In suburban neighborhoods, one distribution transformer serves several houses.

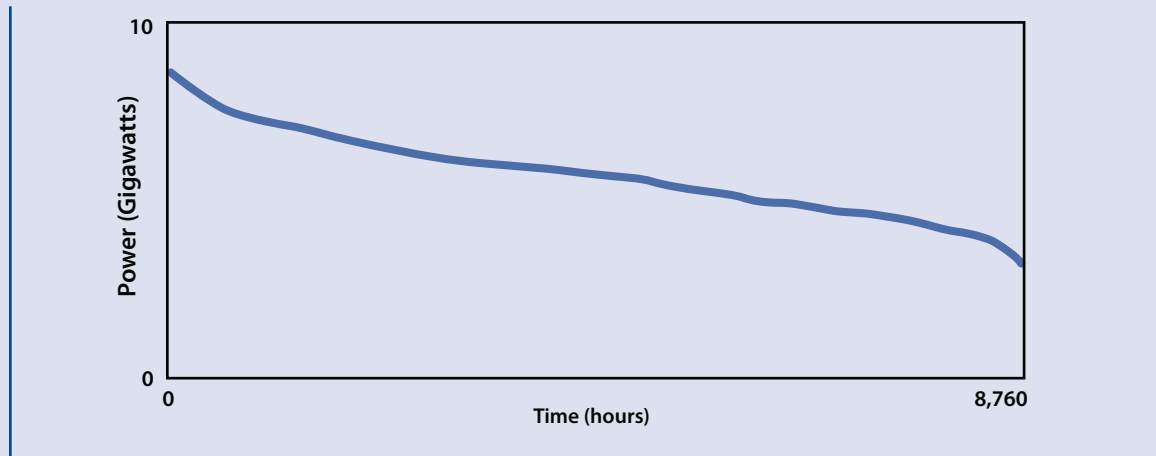
Consumption

Electricity is consumed by a wide variety of loads, including lights, heaters, electronic equipment, household appliances, and motors that drive fans, pumps, and compressors. These loads can be classified based on their impedance, which can be resistive, reactive, or a combination of the two. In theory, loads can be purely reactive, and their reactance can be either inductive or capacitive. However, in practice the impedance of most loads is either purely resistive or a combination of resistive and inductive reactance. Heaters and incandescent lamps have purely resistive impedance, while motors have impedance that is resistive and inductive. Purely resistive loads only consume real power. Loads with inductive impedance also draw reactive power. Loads with capacitive impedance supply reactive power.

Because of the abundance of motors connected to the network, the power system is dominated by inductive loads. Hence, generating units have to supply both real and reactive power. Since capacitors produce reactive power, they often are connected close to large inductive loads to cancel their reactive power (i.e., increase the effective power factor of the load) and reduce the burden on the network and the generators.

From the power system’s operational perspective, the aggregate power demand of the loads in a region is more important than the power consumption of individual loads. This aggregate load is continuously varying. A useful representation of this load across the year is the load duration curve, which plots the load for each hour of the year, not chronologically, but instead by beginning with the hour with the

Figure B.5 A Load Duration Curve



largest load and continuing in a monotonically decreasing fashion, as shown in Figure B.5. For each point on this curve, the horizontal coordinate is the number of hours in the year for which the load is above the power given by the vertical coordinate. The load duration curve provides a good picture of how widely the load varies and for how many hours in a year it is above a particular level. It is more expensive to meet the needs of a spiked load duration curve than a flat one, as generation capacity to meet the peak load is needed, while the generation's utilization is related to the average load. One useful metric of power consumption is the load factor, which is the ratio of average to peak load.

B.4 OPERATION OF THE ELECTRIC POWER SYSTEM

The electric power system is operated through a combination of automated control and actions that require direct human (system operator) intervention. The main challenge in operating the electric power system is that there is negligible "electrical" storage in the system.^{ix} Hence, supply and consumption of electrical power

must be balanced at all times. Since the load is changing all the time in ways that cannot be perfectly predicted, generation must follow the load in real time. The balance between supply and demand is maintained using a hierarchical control scheme, with crude matching at the longer timescale and finer matching at the shortest timescale (see Figure 2.1 in Chapter 2).⁵

Protection

An important aspect of the operation of the electric power system is protection. This means ensuring the safety of the system, including generating units and other grid assets, and the people who may come in contact with the system. Protective action must be taken in fractions of a second to avoid equipment damage and human injury. Protection is achieved using sensing equipment as well as circuit breakers and other types of switches that can disconnect and de-energize parts of the system in the case of a fault, such as a damaged transmission line or a short circuit. Once the fault is repaired, that segment of the system can be brought back online.

^{ix} Note that pumped storage, which uses electricity to pump water into an elevated reservoir and stores energy in the form of potential energy, is not a form of electrical storage. A hydroelectric generating unit must be run to convert this energy back into electrical form. Energy storage technologies are discussed in Chapter 3.

Proactive planning for contingencies also protects the electric power system. Computers are regularly calculating system power flows and voltages under various possible contingencies, for example the failure of a large generator or transmission line, to identify the best corrective action to take in each case.

Real-time Operation

The objective of real-time operation of the electric power system is to ensure that the system remains stable and protected while meeting end user power requirements. This requires a precise balance between power generation and consumption at all times. If this balance is not maintained the system can become unstable—its voltage and frequency can exceed allowable bounds—and result in damaged equipment as well as blackouts. If the balance is not restored sufficiently quickly, a local blackout can grow into a cascading blackout similar to the ones in the U.S. in 1965 and 2003. Fortunately, the stored kinetic energy associated with the inertia of generators and motors connected to the system helps overcome small imbalances in power and this “ride-through” capability gives enough time for an active control system to take corrective action. The balance between supply and demand at the shortest timescale is maintained actively via governor control.

Governor Control

As the load and/or generation changes, altering the balance between demand and supply, the generators on governor control take the first corrective action. The governor is a device that controls the mechanical power driving the generator via the valve limiting the amount of steam, water, or gas flowing to the turbine. The governor acts in response to locally measured changes in the generator’s output frequency from the established system standard, which is 60 Hz in the U.S.^x

If the electrical load on the generator is greater than the mechanical power driving it, the generator maintains power balance by converting some of its kinetic energy into extra output power—but slows down in the process. On the other hand, if the electrical load is less than the mechanical power driving the generator, the generator absorbs the extra energy as kinetic energy and speeds up. This behavior is known as “inertial response.” The frequency of the ac voltage produced by the generator is proportional to its rotational speed. Therefore, changes in generator rotational speed are tracked by the generator’s output frequency. A decreasing frequency is an indication of real power consumption being greater than generation, while an increasing frequency indicates generation exceeding power consumption. Any changes in frequency are sensed within a fraction of a second, and the governor responds within seconds by altering the position of the valve—increasing or reducing the flow to the turbine. If the frequency is decreasing, the valve will be opened further to increase the flow and provide more mechanical power to the turbine, hence increasing the generator’s output power, bringing demand and supply in balance and stabilizing the speed of the generator at this reduced level. The speed of the generator will stay constant at this level as long as the mechanical power driving it balances its electrical load. While very fast, for stability reasons, governor control is not designed to bring the frequency of the generator back to exactly 60 Hz. Correcting this error in frequency is the job of the slower automatic generation control (AGC), discussed later in this section.

Voltage Control

Just as an imbalance in supply and demand of real power causes a change in system frequency, an imbalance in supply and demand of reactive power causes a change in system voltages. If the reactive power consumed by the load increases

^x The generator’s output frequency is proportional to its rotational speed, and traditionally governors have been designed to sense this speed.

without a commensurate increase in reactive power supply, the output voltage of the generator will decrease. Conversely, the output voltage of the generator will increase if the generator is supplying more reactive power than is being drawn. The voltage can be restored to its original level by either adjusting the generator's rotor current (which controls the amount of reactive power produced by the generator), or by using ancillary voltage support equipment, such as static VAR compensators that employ inductors and capacitors in conjunction with semiconductor switches to absorb or supply the imbalance in reactive power. Voltage control is also extremely fast.

Automatic Generation Control

While governor control brings supply of and demand for real power in balance, it results in a small change in system frequency. Furthermore, governor-based reaction of generators located outside a control area to load changes inside the control area (or vice versa) can alter power flows between control areas from their scheduled levels.^{xi} The errors in frequency and flows between control areas are corrected by the relatively slower AGC. AGC aims to eliminate the area control error (ACE). ACE is a measure of both the difference between actual and scheduled net power flows to or from a control area and the error in system frequency. Ignoring the effect of system frequency, a positive ACE means that generation within the area exceeds load by more than the scheduled net power flow from the control area. In this case, the generation in the control area needs to be reduced. Conversely, negative ACE requires local generation to be increased. The area control center automatically sends signals to generators equipped with AGC to increase or decrease their output. In exceptional circumstances, when the required change in output is greater than the defined limit of AGC, the

system operator can call the generation operator over the phone and ask for an increase or decrease in output.

Reserves

Beyond a certain level of power imbalance, system operators need to call in generation reserves. These may be additional generating units that are on standby or generators that are already producing power but can ramp up their output on request. Having adequate reserves on the system is essential to deal with load uncertainties and contingencies, such as the failure of a generating unit.

Reserves are categorized based on the time it takes them to start delivering the requested power; typical categories are 10-minute and 30-minute reserves. Reserves can be either spinning or non-spinning. Spinning reserves are generating units with turbines spinning in synchronicity with the grid's frequency without supplying power. They can deliver the requested power within a few minutes. Non-spinning reserves are units that are offline but also can be synchronized with the grid quickly. In systems with organized markets, reserves are paid not only for the energy they produce but also for being available on short notice to deliver reserve power.

Other Power Balancing Options

Large customers in some regions often face real-time pricing, which induces them to cut loads when the system is under stress and the real-time incremental cost of supplying power is accordingly high. However, when all other options for balancing power have been exhausted, the system operator must resort to proactively reducing the load, generally referred to as load shedding. Load shedding can be accomplished in a number of ways. At first the system operator can interrupt power to those

^{xi} From an operational perspective, a large electric power system is divided into multiple control areas, also called "balancing authority areas." These control areas are connected together via transmission lines that are called "tie-lines."

loads with which they have contracts that permit this. Alternatively, the system operator can order voltage reductions, also known as brownouts. Many loads, such as heaters, incandescent lamps, and certain types of motors, consume less power (and do less work) when operated on a lower voltage. Hence, by reducing the voltage supplied to the loads, the total system power consumption can be reduced. If neither method achieves the desired reduction in load, the system operator can initiate rotating blackouts. In rotating blackouts, groups of consumers are disconnected one at a time in a rotating fashion for a certain fixed duration (typically one hour). This disconnection is typically carried out by opening switches at the distribution substations.

Scheduling

Scheduling determines which generating units should operate and at what power level, and it is accomplished on a predetermined, fixed time interval. The objective is to minimize cost, subject to generation and transmission constraints. Scheduling consists of economic dispatch and unit commitment, each covering two overlapping time ranges.

Economic Dispatch

The incremental production costs of generating units can be quite different from one another, mostly due to differences in the costs of their “fuel” (for example, uranium, coal, natural gas) and their efficiencies. Economic dispatch minimizes overall production costs by optimally allocating projected demand to generating units that are online. Computers at control centers run optimization algorithms, typically every 5 or 10 minutes, to determine the dispatch for the next hour and send these

economic dispatch signals to all the generators. Sometimes power cannot be dispatched from the lowest-cost generating unit due to physical limits of the system or security constraints associated with maintaining secure operation under contingencies. Physical restrictions include transmission lines’ thermal and stability constraints and limitations on generating units’ output power and ramp rates. Security constraints include transmission line reserve capacity and generation reserve requirements. Economic dispatch optimization subject to security constraints is known as “security-constrained economic dispatch.”

Unit Commitment

In addition to determining the amount of power each generating unit should be producing when it is online, system operators must also determine when each generating unit should start up and shut down. This function is known as “unit commitment.” Although significant costs are associated with the startup and shutdown of generating units, it is not practical to keep all of them online all the time. There are large fixed costs associated with running generating units, and some units have a minimum power they must produce when they are online. Unit commitment determines the economically optimal time when generating units should start up and shut down and how much power they should produce while they are online. This optimization is more complex and time consuming than economic dispatch. Unit commitment is typically done one day ahead and covers dispatch for periods ranging from one to seven days.

B.5 WHOLESALE ELECTRICITY MARKETS

The organizational structure of the electric power industry has changed significantly over the last 15 years, as discussed in Chapter 1. Until the mid-1990s, the electric power industry in the U.S. mostly was vertically integrated: a single entity, a regulated monopoly, owned and operated generation, transmission, and distribution in each region.^{xiii} However, in 1996 the Federal Energy Regulatory Commission issued Order No. 888, which required that the transmission network be made available for use by any generator. Since then independent system operators (ISOs) and regional transmission organizations (RTOs) have been created in certain parts of the US. In many regions, ownership of generation and transmission have been separated. In regions where they exist, ISOs and RTOs coordinate organized wholesale electricity markets in which independent decisions of market participants (those who buy and sell energy or other electricity market products, such as spinning reserves) set the price of energy generation, respecting the requirements of central coordination provided by the ISO or RTO.

The theory of spot pricing provides the foundations for successful market design.⁶ In a framework known as “bid-based, security-constrained, economic dispatch,” central coordination by the system operator is integrated with decentralized decisions by market participants. The process of selling wholesale energy begins with a bidding process whereby generators offer an amount of energy for sale during specific periods of the day next day at a specific price. These offers are arranged by the ISO/RTO in ascending order, called the “bid stack,” and the generators are dispatched (told to generate) in this order until generation matches expected load. (Large loads also sometimes submit bids for the purchase of

energy in the market.) All the dispatched generators receive the same compensation, called the “clearing price”—the offer of the last generator dispatched. The actual process is more complicated than this simple explanation, incorporating such parameters as the time required to start generators, out-of-economic-order dispatch due to congestion or reliability concerns, and security constraints. The goal of the system operator is to determine the dispatch that minimizes total cost, as measured by generators’ bids, subject to security constraints.

This process determines the marginal cost of meeting an increment of load at each location (called a “node”) in the transmission system to which load or generation is connected. These costs are termed “locational marginal prices” (LMPs) and are the prices at which transactions for purchasing or selling energy in the market take place. Distribution companies or large customers pay the applicable LMP for energy consumed. Similarly, generation is paid the LMP at the point at which it is located.

The LMP pricing structure used in modern markets ensures that the profitable choice for generators and loads is to follow the instructions of the economic dispatch. Generators are only dispatched when their offer to sell is at a price no greater than the market-clearing price at their location. Likewise, generators are not dispatched when the market price is less than their offer to sell. The use of LMPs allows for the preservation of the traditional industry approach of security-constrained, economic dispatch in the presence of independent system operators and organized wholesale markets. The use of LMPs exploits the natural definition of an efficient equilibrium for a market, utilizes the unavoidable central coordination, and avoids the need for market participants to track transmission flows or understand the many constraints and requirements of the power system.

^{xiii} The exceptions were small municipal and cooperative entities that were distribution-only operations and, particularly from the 1930s on, federal systems such as the Tennessee Valley Authority.

B.6 POWER SYSTEM PLANNING

Construction of new generating units and transmission lines requires large investments and significant time (ranging from a few years to a decade). Hence, planning of electrical power system expansion requires careful analysis that relies on long-term demand forecasts of 10 to 20 years. Projecting demand accurately over the long term is challenging and requires consideration of a number of factors, including estimates of population growth, historic individual consumption patterns, and projected economic growth. Long-term demand forecasts also may incorporate the projected impacts of new energy conservation and demand response programs.

In regions served by vertically integrated utilities, generation and transmission expansion planning is carried out centrally by system planners at the utilities. Planners evaluate various options for meeting future load demand in terms of capital and operating costs. They select projects based on minimizing system cost while providing adequately reliable service. Decisions also may be influenced by government incentives, regulations, and environmental impact restrictions. Planning has to allow for the risk associated with the significant uncertainty in long-term load forecasts, future operating costs (directly related to fuel prices), and technological changes.

In regions with organized wholesale markets, expansion planning is split between ISOs/RTOs and the individual market players. Generation planning is decentralized and primarily accomplished by individual generation companies based on forecasts and system needs from the RTO/ISO. Transmission planning is still mostly centralized and coordinated by ISOs/RTOs. While the precise mechanism by which expansion projects are selected depends on market design details, individual company decisions are based on maximizing return on investment. However, the competitive nature of the market is expected to lead to an overall system cost minimization while providing stronger incentives for operating efficiencies.

An added complexity in planning future transmission expansion in areas with organized markets is the uncertainty associated with future generation investments. Transmission expansion decisions are more challenging for ISOs/RTOs because development of generating plants is based on individual company decisions. As a result, the ISOs/RTOs cannot know the location and size of these future plants with certainty.

REFERENCES

- ¹U.S. Energy Information Administration, “Electricity Explained: Use of Electricity,” http://205.254.135.24/energyexplained/index.cfm?page=electricity_use.
- ²W. Steinhurst, “The Electric Industry at a Glance” (Silver Spring, MD: National Regulatory Research Institute, 2008).
- ³S. W. Blume, *Electric Power System Basics: For the Nonelectrical Professional* (Hoboken, NJ: Wiley–IEEE Press, 2007).
- ⁴A. V. Meier, *Electric Power Systems: A Conceptual Introduction* (Hoboken, NJ: Wiley–IEEE Press, 2006).
- ⁵I. J. Pérez-Arriaga, H. Rudnick, and M. Rivier, “Electric Energy Systems. An Overview,” in *Electric Energy Systems: Analysis and Operation*, eds. A. Gomez-Exposito, A. J. Conejo, and C. Canizares (Boca Raton, FL: CRC Press, 2009), 60.
- ⁶F. Schweppe, M. C. Caramanis, R. D. Tabors, and R. E. Bohn, *Spot Pricing of Electricity* (Boston, MA: Kluwer Academic Publishers, 1988).