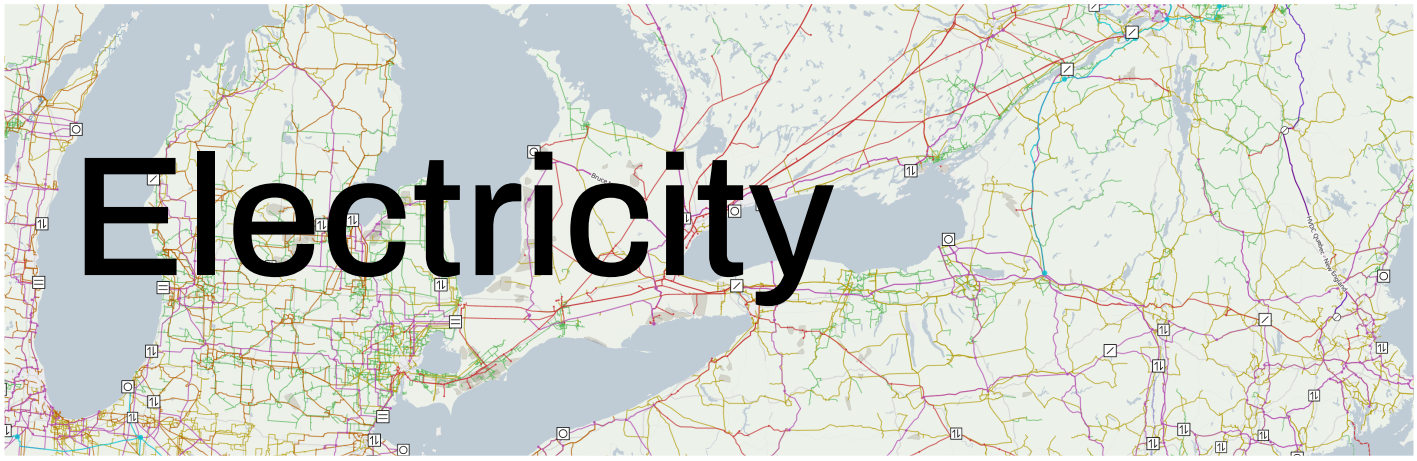
How



Electricity Markets

Work

KEEGAN TRUJILLO-GREEN

1 Introduction

The world's electrical grids stand alongside the internet and water supply networks as great systems and accomplishments of humanity. We rely on these systems every day, yet most of us do not understand how these complex systems work. We often take these systems for granted when they are functioning correctly, which they almost always do. When such systems fail, such as during a power outage, we realize how dramatically they improve our quality of life.

The ability to flip a switch and flood a room with light in the middle of the night is the result of over a hundred years of human innovation and people working around the clock to keep the world's grids running smoothly—to literally keep the lights on. What happens every time we turn on an appliance often eludes us. Nonetheless, without fail, slightly more current gets drawn through tens or hundreds of kilometers of transmission and distribution lines. And, in tandem, the generators at the end of those transmission lines start to work slightly harder, whether in a power station or a solar or wind farm.

But even this is not a full appreciation for what goes into constantly balancing the supply and demand of electricity. The grid is just as much an economic system as a physical one, and it can be operated in a way that creates the perfect economic playing field—one that incentivizes the generation of power that is plentiful, reliable, and affordable. That playing field is the wholesale electricity market. But before we discuss the electricity market, let's look at a brief history of the grid.

1.1 A Brief History of the Electrical Grid

The electric power industry was pioneered in the US and UK in the late 1800s and early 1900s [Tuttle, 2016]. Before electrical grids were established, the supply and demand of electricity had to be co-located. This was inconvenient and inefficient, so private utility companies came about to operate their own generation and distribution systems serving multiple customers [Borberly, 2001]. In some cases, utility companies connected their respective systems to be able to work together to meet peaks in demand. It was quickly realized that electricity systems serve the public good, and government oversight resulted in a new world order for utility companies. Around 1940 in the UK and France, nationalization resulted in the interconnection of regional grids into nationwide grids [NT, Carrive]. And by 1960, many European countries had nationwide grids of one form or another [Schmalensee, 2021].

Whether owned by a private company or government organization, and whether the size of a region or of a country, electrical grids must be operated such that the supply of electricity equals the demand for electricity at all times. Otherwise, as we will learn, load shedding, brownouts, or blackouts occur. Enter the wholesale electricity market.

1.2 Wholesale Electricity Markets

A wholesale electricity market, also known as a power exchange (PX), is a system that allows balancing supply and demand while minimizing the cost of electricity production (supply) and maximizing value to electricity consumers (demand). In a wholesale electricity market, loads participate by *bidding* to buy power, and generators participate by *offering* to sell power. A generator may be a power station (e.g., a hydroelectric power station) or a wind or solar farm. A load may be a utility company (participating in the market on behalf of its customers) or, more recently, large industrial consumers.

A central authority decides which bids and offers to accept/dispatch, or reject. Making this decision is called *clearing the market*, the objective of which is to achieve the best overall outcome for the generators and loads. Because of this objective, the market clearing process has some key requirements. Firstly, the decision must be informed by knowledge of the transmission and distribution networks, of all generators and their offers, and of all loads and their bids. The central authority must know the limits of the grid's transmission lines, otherwise they might try to transmit too much power through certain transmission lines, rather than generating locally when necessary. The second key requirement of the market clearing process is that it must be done for all generators and loads simultaneously, as part of one decision, as there is only one outcome that maximizes benefit to all generators and loads.

The only way these requirements can be met is if the decision is made by a central authority, rather than by any one generation company or utility. The central authority is typically responsible not only for clearing the market, but for operating the market in general. The central authority is usually a government-owned or government-appointed not-for-profit organization, but sometimes, especially historically and in the US, it is a privately owned and for-profit company. However, the central authority being a government organization—one that does not itself own (or have stakes in) any generators or loads—ensures that the operation of the market is unbiased. Because such an organization cannot itself earn rev-

enue from generating electricity, it operates for the public good, not for profit. Thus, generators are paid fairly, loads are charged fairly, and both are selected for dispatch fairly. In North America, such organizations are called Independent System Operators (ISOs) because they do not own any—they are independent of—generators, loads, and transmission lines. Examples include the California ISO (CAISO) and Ontario's Independent Electricity System Operator (IESO).

1.3 Open Versus Closed Markets

We previously discussed private companies versus government organizations as the central authority for operating an electricity market. We discussed that, in the case of a government organization, it should not have stakes in any generators or utility companies. However, this says nothing of whether generators or utility companies should themselves be private or government-owned. In fact, as part of energy liberalization trends occurring at the turn of the millennium, many governments around the world opened their electricity markets to allow more private generation companies to participate, rather than just government-owned or government-contracted ones. This attempted to reduce electricity prices by increasing competition. Such energy liberalization efforts are thus also known as *privatization* or market *deregulation*, although the latter term can be misleading as it may be taken to mean that government regulations surrounding electricity market operations were relaxed, which is certainly not the case. Energy liberalization often came with restructuring of the government organizations responsible for the grid. For example, IESO, like many ISOs, was founded in the late 1990s, but that does not mean that Ontario did not have an electricity market before then, merely that it was closed off to participants that were not government-owned or government-contracted, and thus less competitive.

2 Supply and Demand of Electricity

2.1 Introduction to Economics: Your Friendly Neighborhood Espresso Market

Imagine a neighborhood of cafés, all serving espresso. What a wondrous place to live! The baristas make as many espresso shots as customers buy. More specifically, at any given time, the rate at which baristas supply espresso shots (e.g., in shots per hour) is always equal to the rate at which customers demand them.

For reasons that will become apparent later, we will assume that all cafés serve identical espresso and offer an identical experience to customers. Because the cafés' offerings are the same and they are conveniently located within the same neighborhood, the only thing that would set them apart in the eyes of customers is their prices.

But if a café tries to undercut its neighbors, the other cafés will follow suit if the lower price is sustainable in order to stay competitive. And if a café raises its price to increase its profit margin, others will be able to do the same. For a given time, place, and offering, there is one stable price. Different cafés may have different operating costs—perhaps a café has a more expensive supplier of espresso beans—but they tend to charge the same price per espresso shot and pocket the difference.

Customers' demand for espresso shots is based on how much the cafés charge for each one; there will be fewer customers if the cafés charge more and vice versa. Thus, the rate at which customers buy espresso is a function of the price per espresso shot. This function is called the demand curve, which models buyer behavior in aggregate and is generally monotonically decreasing. The demand curve changes with time; for example, demand for espresso is greatest in the morning.

If the cafés have to make more espresso, they will have to purchase more espresso beans. They will have to call upon additional baristas to help, or in general have to pay each barista more per hour—more per shot, essentially. In other words, the cafés have a cost per espresso shot, which is reflected in the price they charge for each one. However, the cafés' price per shot is not constant. If the cafés are overwhelmed with customers, they may charge more for each espresso shot to take advantage of the situation and make more profit per shot. The cafés may have to augment their baristas' hourly wages with bonus pay to keep them happy. In order to meet the demand, the cafés may also have to bust out their older, less efficient espresso machines from their back rooms that use more espresso beans per shot. On the other hand, if the cafés are having a slow day, they may reduce their price and sacrifice some profit per espresso shot in order to make the most of the low demand. This is called dynamic pricing. The price is a function of the rate at which espresso shots are being sold. This function is called the supply curve, which models seller behavior in aggregate and is generally monotonically increasing.

A given café is willing to sell espresso at or above a certain price, and a given customer is willing to buy at or below a certain price.

2.2 The Market Clearing Process

The above principles are true not just for a neighborhood of cafés but for any capitalist market. These basic tenets of economics are the principles of supply and demand. They apply wherever a quantity Q of a particular good is exchanged at a price P. Q is measured in goods per hour, and P is measured in dollars (or any other currency) per good.

The allocation of buyers to sellers (customers to cafés) is a process called *clearing the market*. This determines the quantity of goods that gets exchanged in the market, Q^*, and at what price, P^*. This price is known as the *market price* or *market-clearing price*. Sellers who charge a low price are able to produce at a low cost and do not price-gouge. A capitalist market rewards such sellers, giving them priority over other, less efficient ones. At the same time, a capitalist market rewards buyers who are willing to pay a high price, giving them priority over other buyers who don't value the product as much. It follows that we can model the market clearing process by allocating an increasing quantity of goods to the buyers and sellers, starting with the buyers who are willing to pay the most and the sellers who are willing to charge the least, until the buying price equals the selling price.

We can represent this model graphically (or mathematically) by plotting the supply and demand curves as shown in Figure 1. The market-clearing price P^* and quantity Q^* are dictated by the point where the two curves intersect. The area under the supply curve until a quantity Q represents the cost C of producing that quantity. The area under the demand curve until a quantity Q represents the cumulative value delivered to buyers by that quantity, known in economics as *utility*, U. Subject to the constraint that the quantity sold must be the same as the quantity bought, Q^* simultaneously maximizes buyer utility and minimizes seller cost. The difference $U - C$ is known in economics as *surplus* or *welfare*, W. In clearing the market, welfare is maximized.

Welfare can be split into two parts: producer surplus, and consumer surplus. Producer surplus is the area above the supply curve but below P^*, representing the profit earned by selling above the cost of production. Consumer surplus is the area below the demand curve but above P^*, representing the extra value delivered to customers beyond what they paid for.

Obviously there is no regulatory body that allocates customers to cafés, no economist planted in town square finding the intersection between the supply and demand curves. Your friendly neighborhood espresso market clears itself. Cafés try to charge the highest price that the optimal number of customers are willing to pay, and customers naturally seek out the lowest price that cafés are willing to charge; the two sides generally meet somewhere in the middle.

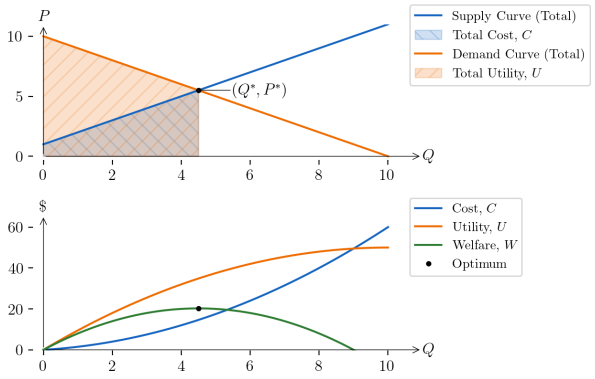


Figure 1: Supply and demand curves.

2.3 The Electricity Market

The principles of supply and demand are so universal that they apply to the power grid when operated as an electricity market. In this case, the good is not espresso shots but megawatt-hours of electricity—energy for machines, rather than for humans. The unit of measurement is not espresso shots per hour but megawatt-hours per hour—so, just megawatts (MW). In an electricity market, the rate at which electricity is supplied (the production, or generation) is always equal to the rate at which it is demanded (the consumption, or load). In the café analogy, generators are espresso machines, and loads—such as your local utility company—are customers. Everything that we've discussed so far in terms of cafés (and capitalist markets in general) applies to the electricity market.

We assumed before that all cafés served identical espresso. In other words, we assumed that espresso is a *commodity*—a good that varies little between sellers and that buyers need not be picky about. Obviously this isn't true about espresso in the real world! Examples of commodities include wheat, crude oil, and...electricity! Indeed, all electrons are created equal. Electricity may be produced from a variety of energy sources (nuclear, hydro, and so on) and travel varying distances to get to your home, but there is no differentiating between a nuclear-generated electron and a hydro-generated electron. People care about the origin and locality of their food more than that of their electricity. Yes, some people care about the environmental sustainability of their electricity—

I among them—but we cannot choose between sustainable and unsustainable electrons. We can only choose to consume electricity at times of day when there is a higher proportion of sustainable electrons in the mix. But, at a given time, all electrons are equal.

Previously we discussed the market clearing process. The electricity market must be cleared like any other, but unlike a neighborhood espresso market, it cannot feasibly clear itself. There are two reasons for this.

First of all, there is no buffer between production and consumption. If cafés are busy, customers can feasibly wait for up to several minutes between placing their orders and receiving their espresso. Alternately, the cafés can prepare extra espresso shots while morning demand is still ramping up, to reduce customer wait times. But this does not work in the electricity market. When you turn on your kettle, you don't have to wait even a second for power. You just take. The grid cannot ramp up in advance of everyone turning on their kettles in the morning because, traditionally, there is no way to store that extra power.

The second reason for which the electricity market cannot clear itself is that resources are pooled. When you want espresso, you go to one café and talk to one barista who fulfills your order. But when you want to turn on your electric kettle at home, you don't call up a generator and ask them to ramp up by a couple kilowatts. There is a disconnect between the demand (you, or rather, your utility company) and the supply (the generators).

Thus, the electricity market must be administered by a regulatory body—the ISO—that monitors the demand, and tells each generator how much to supply such that the total supply exactly equals the demand at all times. It is worth it to have an organization that is different to the company or companies who own and operate the generators, because it fosters healthy competition between the generation companies, which reduces the cost of electricity.

2.4 Bids to Buy and Offers to Sell Power

Loads and generators submit bids to buy power and offers to sell power, respectively. Bids and offers are known in general as market submissions. Each submission includes a list of price-quantity pairs. In a given price-quantity pair, the quantity is the maximum quantity of power that a generator is willing to sell, or that a load is willing to buy, at the corresponding price in the pair. The price is measured in dollars per MWh, and the quantity is measured in MW. For an offer, the price must monotonically increase as the quantity increases, and for a bid, the price must monotonically decrease as the quantity increases.

If these sound similar to supply and demand curves, it's because they are! In an electricity market with one generator and one load, the generator's (P_G, Q_G) pairs would define the market's supply curve and the load's (P_L, Q_L) pairs would define the market's demand curve. The two curves' intersection determines the quantity Q^* of power that will be delivered at the market-clearing price P^*, which maximizes $U - C$.

2.4.1 Example 2.1: One Generator, One Load

Consider the example shown in Figure 2. The generator says they are willing to produce up to 6 MW at \$2/MWh, or up to 9 MW at \$7/MWh. Note that this does not mean they are willing to produce the first 6 MW at \$2/MWh and only the remaining 3 MW at \$7/MWh; the generator has one price, and it changes if a certain volume is surpassed, like in the café analogy. The same principle applies on the demand side: the load says they are willing to consume up to 4 MW at \$8/MWh, or up to 8 MW at \$5/MWh. The market clears at $Q^* = 6$ MW and $P^* = \$5$/MWh.

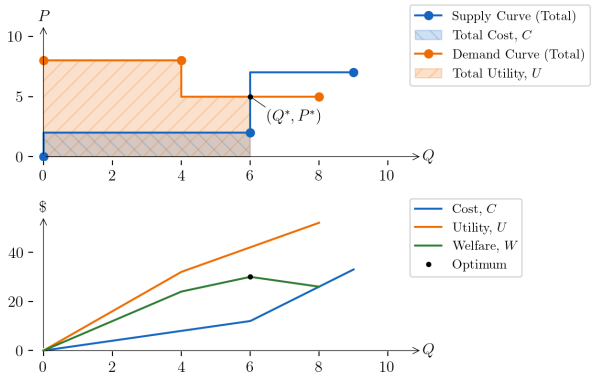


Figure 2: Supply and demand curves formed by the price-quantity pairs of the one generator and one load of Example 2.1.

2.4.2 Example 2.2: Two Generators, Two Loads

What if there are two generators, G1 and G2, and two loads, L1 and L2? Each generator and load has its own supply or demand curve formed by its price-quantity pairs, as shown in Figure 3. These are known as individual supply and demand curves, as they represent individual participants in the market.

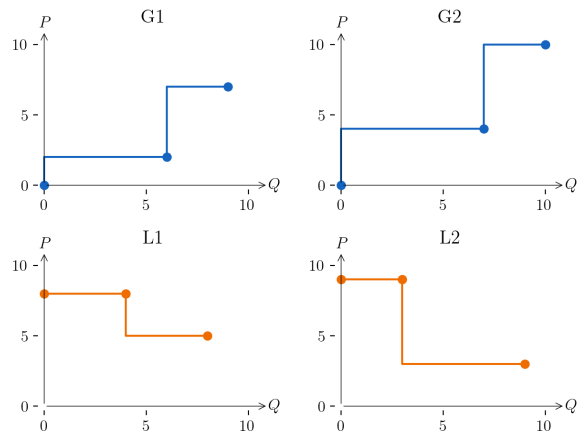


Figure 3: Individual supply and demand curves formed by the price-quantity pairs of the two generators and two loads of Example 2.2.

In clearing the market, we still want to maximize $W = U - C$ subject to the constraint that the total quantity sold, $Q_{G1} + Q_{G2}$, must equal the total quantity bought, $Q_{L1} + Q_{L2}$. This constraint represents conservation of power flow.

$$\begin{aligned} &\max_{Q_{G1}, Q_{G2}, Q_{L1}, Q_{L2}} W(Q_{G1}, Q_{G2}, Q_{L1}, Q_{L2}) \\ &\text{subject to} \quad Q_{G1} + Q_{G2} = Q_{L1} + Q_{L2} \end{aligned} \tag{1}$$

Where the total seller cost is the total area under each generator's individual supply curve, up to the quantity produced by each generator:

$$C(Q_{G1}, Q_{G2}) = \int_0^{Q_{G1}} p_{G1}(q)\,dq + \int_0^{Q_{G2}} p_{G2}(q)\,dq \tag{2}$$

And the total buyer utility is the total area under each load's individual demand curve, up to the quantity consumed by each load:

$$U(Q_{L1}, Q_{L2}) = \int_0^{Q_{L1}} p_{L1}(q)\,dq + \int_0^{Q_{L2}} p_{L2}(q)\,dq \tag{3}$$

This optimization problem can be simplified by splicing the generators' and loads' individual supply and demand curves into an aggregated supply curve and an aggregated demand curve, respectively. The optimization problem can then be solved by finding the intersection between the aggregated supply and demand curves. The solution to the example at hand is $Q^* = 11$, $P^* = 4$. This is shown in Figure 5, and Figure 4 shows how Q^* corresponds to the individual quantities Q^*_{G1}, Q^*_{G2}, Q^*_{L1}, and Q^*_{L2}. The optimal cost, utility, and welfare are tabulated in Table 1.

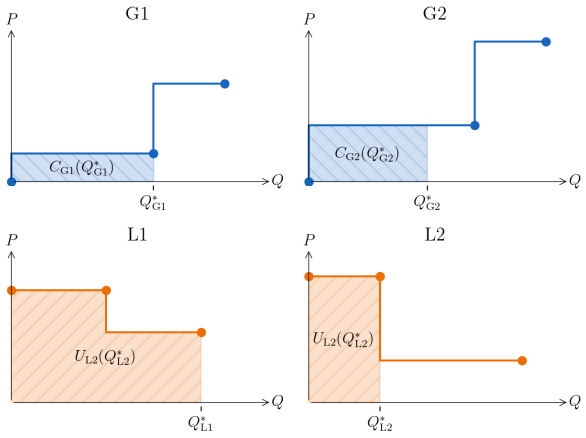


Figure 4: Individual supply and demand curves of the two generators and two loads of Example 2.2, showing their optimal costs and utility values, respectively.

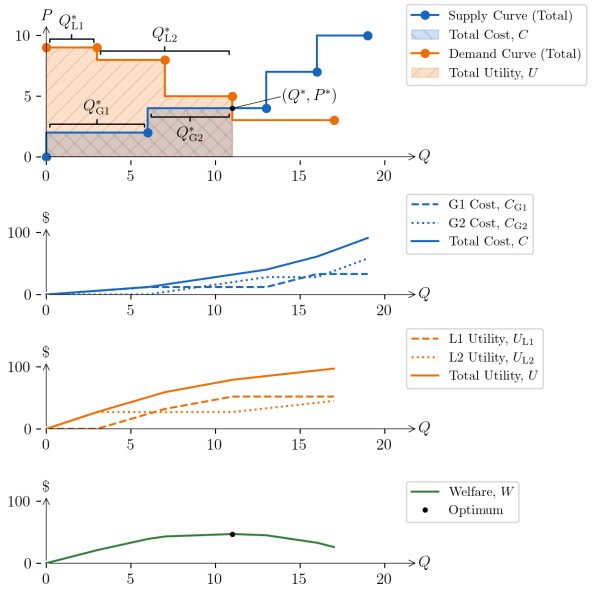


Figure 5: Aggregated supply and demand curves of the two generators and two loads of Example 2.2, respectively, showing the resulting market-clearing price and quantity.

2.5 The Market Network

We have established that for a given market, at a given time, there is one market price. For the purposes of market modeling, the grid consists of multiple pricing nodes, each of which has its own market

	Cost	Utility	Producer Surplus	Consumer Surplus	Welfare
	$C_{G1} =$ \$12.0/h	$C_{L1} =$ \$52.0/h	$S_{G1} =$ \$12.0/h	$S_{L1} =$ \$15.0/h	
	$C_{G2} =$ \$20.0/h	$C_{L2} =$ \$27.0/h	$S_{G2} =$ \$0.0/h	$S_{L2} =$ \$20.0/h	
Total:	\$20.0/h	\$79.0/h	\$12.0/h	\$35.0/h	\$47.0/h

Table 1: Optimal cost, utility, and welfare of Example 2.2.

	Cost	Utility	Producer Surplus	Consumer Surplus	Welfare
	$C_{G1} =$ \$12.0/h	$C_{L1} =$ \$32.0/h	$S_{G1} =$ \$12.0/h	$S_{L1} =$ \$11.0/h	
	$C_{G2} =$ \$13.$\bar{3}$/h	$C_{L2} =$ \$27.0/h	$S_{G2} =$ \$0.0/h	$S_{L2} =$ \$10.$\bar{6}$/h	
Total:	\$25.$\bar{3}$/h	\$59.0/h	\$12.0/h	\$21.$\bar{6}$/h	\$33.$\bar{6}$/h

Table 2: Optimal cost, utility, and welfare of Example 2.3.

price and in that sense acts like a market. Each node is analogous to one neighborhood of cafés. But, of course, the grid is an electrical network, and power can be transferred between its nodes through long-distance transmission and distribution lines. This is as if the neighborhoods of cafés were interconnected by an intricate network of pipes! Like electricity, espresso wouldn't even have to be made in the neighborhood in which it is consumed, and supply does not need to equal demand within a grid node. Because quantities of power can be traded between grid nodes, the entire grid must be cleared as one market in order to maximize the system-wide welfare W. The total welfare is generally higher with trade than without, which is why trade takes place.

However, the network is leaky. Transmission and distribution lines are not ideal wires and have electrical resistance depending on their type and length. Thus, some power is lost to heat for each unit of power transferred. These transmission losses are analogous to a cost of trade between grid nodes. Power transfer between nodes is still usually necessary and beneficial. For example, it is likely be cheaper to use nuclear power transmitted over a great distance, with some lost along the way, than to use power from a nearby gas-fired plant. We need a way to model and account for the costs of trade when clearing the market, as they will affect the optimal quantities of power and cause prices to differ between nodes. In general, each node affects all other nodes.

A line's transmission losses can be modeled by the line's efficiency, η:

$$\eta = \frac{\text{power out}}{\text{power in}} = \frac{\text{power out}}{\text{power out} + \text{power lost}} \tag{4}$$

2.5.1 Example 2.3: One Generator Node, One Load Node

Recall the previous section's example of two generators and two loads. Consider now that the generators G1 and G2 are located at a node A, and that the two loads L1 and L2 are located at a node B. The two nodes are connected by a transmission line with 75% efficiency. This is an unrealistically poor efficiency, for illustrative purposes. The optimization problem becomes maximizing W subject to the constraint that the total quantity bought, Q_L, must equal 75% of the total quantity sold, Q_G, meaning that for every 1 MW generated, only 0.75 MW are delivered to the load, the rest being lost to heat. Of the two quantities Q_G and Q_L, either can be chosen as the optimization problem's sole decision variable, from which the other quantity can be calculated via the efficiency. The solution to the example at hand is $Q_G^* = 9.\bar{3}$ MW, $Q_L^* = 7$ MW, $P_G^* =$ \$4/MWh, $P_L^* =$ \$5.$\bar{3}$/MWh. This is shown in Figure 6. The optimal cost, utility, and welfare are tabulated in Table 2.

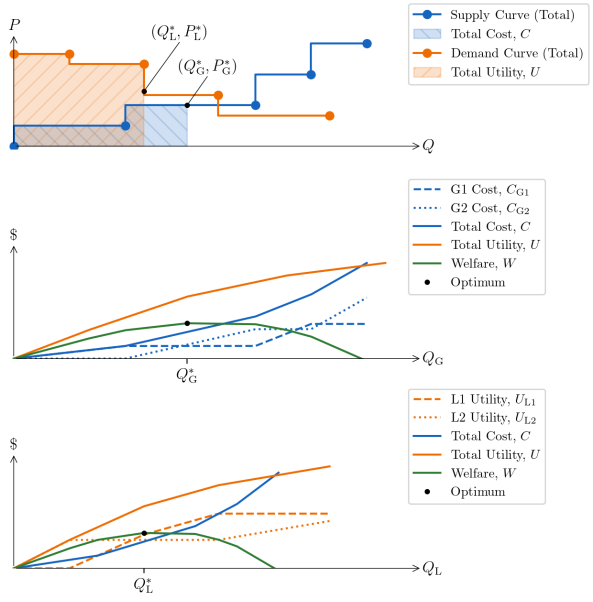


Figure 6: Determining the market-clearing prices and quantities for the two-node grid of Example 2.3.

Notice how the welfare has decreased from \$47.0/h in Example 2.2 to \$33.$\bar{6}$/h in this example. The \$13.$\bar{3}$/h difference is the cost of having to transfer power between the generator and load nodes via lossy transmission lines. In this example, the producer surplus happened to stay the same (\$12.0/h) while the

consumers' surpluses decreased with the introduction of transmission losses. So, in this example, only the consumers incurred the cost of trade. However, in general, the cost of trade is paid by both producers and consumers. This does not mean that trade is not worthwhile. On the contrary—in this example, transmission lines allowed the generator and load to find each other, and trade in general provides more optimal buying and selling opportunities for market participants.

Note
With transmission losses, maximizing $U - C$ is only the same as maximizing $C_L + C_G$ if an additional constraint $Q_L P_L = Q_G P_G$ is applied.

2.5.2 Example 2.4: One Generator Node, One Combined Generator/Load Node

Now consider that only generator G1 is located at node A, while generator G2 and both loads are located at node B. The optimization problem becomes maximizing W subject to the constraint that $0.75\,Q_{G1} + Q_{G2} = Q_L$. Any two of these three quantities can be chosen as the decision variables; we will use Q_{G2} and Q_L arbitrarily. The optimization problem is then:

$$\begin{aligned} &\max_{Q_{G2},\,Q_L} \; W(Q_{G2}, Q_L) \\ &\text{subject to} \quad 0.75\,Q_{G1} + Q_{G2} = Q_L \end{aligned} \tag{5}$$

We can visualize this optimization problem as the 3D surface shown in Figure 7. The parallelogram shape of the surface as viewed from above in Figure 7(b) is determined by the above constraint and by the bounds $0 < Q_{G2} < 10$ MWh.

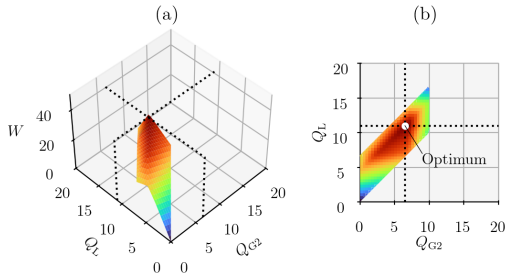


Figure 7: Determining the market-clearing quantities Q_{G2} and Q_L that maximize W in Example 2.4.

The solution to the example at hand is $Q^*_{G2} = 6.5$ MW, $Q^*_L = 11$ MW, and $Q^*_{G1} = 6$ MW. The market-clearing prices at the two nodes are $P^*_A =$ \$3/MWh and $P^*_B =$ \$4/MWh. This is shown in Figure 8.

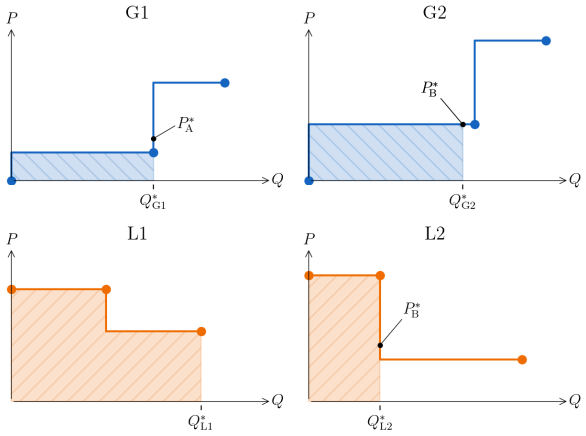


Figure 8: Individual supply and demand curves of generator G1 (at node A) and generator G2, load L1, and load L2 (at node B), showing the resulting market-clearing prices and quantities.

2.6 Balancing Power Supply and Demand

We previously mentioned that supply always equals demand in the electricity market. Indeed, according to the law of conservation of energy, the power flowing into the grid equals the power flowing out of the grid at any given time. However, we just said that an ISO is responsible for balancing power supply and demand in the grid. Furthermore, if you are already familiar with the power grid and electricity market, you will often hear about *balance* or *imbalance* between power supply and demand, even though an imbalance between power in and power out is not technically possible. So, what do people really mean when they talk about imbalance between supply and demand in the grid? To answer this question, we must first understand some nuances of how the grid works.

The power transferred in the grid is driven by AC voltage, which oscillates approximately 50 or 60 times a second depending on the country. This rate of oscillation is called the grid's frequency, and is measured in hertz (Hz). A lot of equipment connected to the grid requires a frequency very close to 50 or 60 Hz.

Most generators in the grid are synchronous electric generators spun by a turbine—a water turbine in a hydroelectric power station, a gas turbine in a gas power station, or a steam turbine in a nuclear or coal power station. The turbine must be fed with water, steam, or burning gas at a certain rate in order

to produce a certain power output via the generator. Furthermore, a synchronous generator must be spun at a certain speed in order to generate AC voltage at the required frequency. A synchronous generator naturally rotates at a speed corresponding to the frequency of the grid to which it is connected. It will be important to know that a synchronous generator, like any moving object, inherently stores some kinetic energy while in operation, like a flywheel.

Consider the situation where one of the grid's generators has just turned off, or where more load is suddenly connected to the grid. More power must instantly be supplied to the grid such that supply continues to exactly meet demand. At first, this extra power is supplied by the grid's synchronous generators—but *not* by consuming more water, steam, or gas. Instead, the grid's synchronous generators will start slowing down at a steady rate, releasing their inherent kinetic energy and converting it to the extra power required. Synchronous generators do this immediately and naturally, due to the laws of physics rather than human intervention. This behavior of synchronous generators is called *inertial response* and applies just as well if load is suddenly disconnected from the grid, in which case the grid's synchronous generators will start speeding up to absorb the sudden surplus of power. The surplus powers the generator like a motor.

You may be wondering what happens when the generator spins down to zero. Don't bother. There would be a blackout well before that occurs. The inertial response helps to balance supply and demand, but at a cost. As the grid's generators speed up or slow down to absorb to release power as required, the grid's frequency increases or decreases. However, grid-connected equipment often requires a grid frequency within a very specific range. The inertial response only grants the ISO a few precious seconds in which to properly adjust power generation by adjusting the rate at which water, steam, or gas is consumed. If this rate cannot be increased because all the grid's generators are already operating at maximum power, then the ISO must curtail or shed load.

In the café analogy, customers are picky. The inertial response would be equivalent to baristas re-using coffee grounds to save time while serving a sudden influx of customers. Disgusting! Cafés must handle an influx of customers by paying their baristas more in order to work faster—i.e., by operating synchronous generators at a higher power setting—or by utilizing additional baristas and espresso machines—i.e., by turning on additional generators. If the cafés are already pushed to their limit, they must turn some customers away rather than continue to serve terrible-quality espresso.

So, it is appropriate (if not slightly misleading terminology) to talk about balance or imbalance between power supply and demand. What people really mean when they say that supply is not meeting electricity market demand is that the power being generated *apart from the inertial response* is not meeting demand. Discussing imbalance is the same as discussing under- or over-frequency.

References

Lighting by electricity. URL https://web.archive.org/web/20110629091025/http://www.nationaltrust.org.uk/main/w-chl/w-places_collections/w-collections-main/w-collections-highlights/w-collections-lighting-electricity.html.

J. F. Borberly, A.; Kreider. *Distributed Generation: The Power Paradigm for the New Millennium*. CRC Press, Boca Raton, FL, 2001.

P. Carrive. Réseaux de distribution – Structure et planification. *Techniques de l'ingénieur*, D4210:6.

R. Schmalensee. *Handbook on Electricity Markets*. Edward Elgar Publishing, Cheltenham, UK, 2021. ISBN 9781788979955. doi: 10.4337/9781788979955.00008.

Tuttle. The History and Evolution of the U.S. Electricity Industry. techreport, The University of Texas at Austin, Austin, TX, 2016. URL https://energy.utexas.edu/sites/default/files/UTAustin_FCe_History_2016.pdf.