Power System Economics

Designing Markets for Electricity

Steven Stoft

IEEE Press
&
WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

ISBN 0-471-15040-1

Contents in Brief

Chapter 1-3

Pricing Power, Energy, and Capacity

It is not too much to expect that our children will enjoy in their homes electricity too cheap to meter.

Lewis L. Strauss
Chairman, Atomic Energy Commission
1954

POWER IS THE RATE OF FLOW OF ENERGY. Similarly, generating capacity, the ability to produce power is itself a flow. A megawatt (MW) of capacity is worth little if it lasts only a minute just as a MW of power delivered for only a minute is worth little. But a MW of power or capacity that flows for a year is quite valuable.

The price of both power and energy can be measured in $/MWh, and since capacity is a flow like power and measured in MW, like power, it is priced like power, in $/MWh. Many find this confusing, but an examination of screening curves shows that this is traditional (as well as necessary). Since fixed costs are mainly the cost of capacity they are measured in $/MWh and can be added to variable costs to find total cost in $/MWh.

When generation cost data are presented, capacity cost is usually stated in $/kW. This is the cost of the flow of capacity produced by a generator over its lifetime, so the true (but unstated) units are $/kW-lifetime. This cost provides useful information but only for the purpose of finding fixed costs that can be expressed in $/MWh. No other useful economic computation can be performed with the "overnight" cost of capacity given in $/kW because they cannot be compared with other costs until "levelized." While the U.S. Department of Energy sometimes computes these economically useful (levelized) fixed costs, it never publishes them. Instead it combines them with variable costs and reports total levelized energy costs.[1] This is the result of a widespread lack of understanding of the nature of capacity costs.

Confusion over units causes too many different units to be used, and this requires unnecessary and sometimes impossible conversions. This chapter shows how to make almost all relevant economic calculations by expressing almost all prices and

1. In Tables 14 through 17 of one such report (DOE 1998a) the useful (amortized) fixed costs are not reported, and the fixed O&M costs are reported in $/kW which may be an amortized value reported with the wrong units or, if the units are correct, may represent a misguided conversion of an amortized cost to an "overnight" cost.

costs in dollars per megawatt-hour ($/MWh). The remainder of the book confirms this by working every example in these units.

Chapter Summary 1-3: Energy is measured in MWh, while power and capacity are measured in MW. All three are priced in $/MWh, as are fixed and variable costs. Other units with the same dimensions (money divided by energy) may be used, but this book will use only $/MWh. Screening curves plot average cost as a function of capacity factor. The slope of the curve is **variable cost**, and the intercept is **fixed cost**. The average cost (AC_K) plotted in these graphs is not the average cost of using a megawatt-hour of energy produced at a certain capacity factor but rather the average cost of a megawatt-hour of generating capacity. Because the equation for a screening curve is used through the book, understanding this distinction is crucial.

Working Summary

Readers wishing to gain only a working knowledge of measurement units for use in later chapters should understand the following.

Quantity	Quantity units	Price Units
Energy	MWh	$/MWh
Power	MW	$/MWh
Capacity	MW	$/MWh

Cost	Symbol	Cost Units
Fixed	FC	$/MWh
Variable	VC	$/MWh
Average	$AC_K = FC + cf \times VC$	$/MWh
Average	$AC_E = FC / cf + VC$	$/MWh

Ratio	Symbol	Units
Capacity factor	cf	none
Duration	D	none

Notes: Energy is a static amount while power and capacity are rates of flow. The average cost of using capacity, AC_K, depends on the capacity factor, cf, which is the fraction of time the capacity is used. The average cost of energy, AC_E, produced by a specific generator also depends on cf.

Section 1: Measuring Power and Energy. Power is the flow of energy and is measured in watts (W), kilowatts (kW), megawatts (MW), or gigawatts (GW). Energy is an accumulation of power over a period of time. For instance, a kilowatt flowing for one hour delivers a kilowatt-hour (kWh) of energy. The price of both energy and power is expressed in $/MWh. It can also be expressed in "mills," short for "milli-dollars per kilowatt hour," with 1 mill equal to $1/MWh.

Section 2: Measuring Capacity. Capacity is the *potential* to deliver power and is measured in megawatts. Like power, it is a flow.

Section 3: Pricing Capacity. "Overnight" capacity costs are measured in $/kW and so cannot be added to or averaged with variable costs to find which generator could more cheaply serve load of a specific duration. Screening curves plot the annual revenue requirement (ARR) of a generator as a function of the generation's capacity factor. Fixed cost (FC) is the value of ARR for a capacity factor of zero. Since ARR is measured in $/kWy, the same must be true of fixed cost. Dividing FC by 8.76 converts it to $/MWh, a more convenient set of units. Considering the rental cost of capacity makes these units seem more natural.

To avoid confusion when using screening curves and their associated algebra, the distinction between the average cost of capacity (AC_K) and the average cost of energy (AC_E) should be kept in mind. Traditional screening curves graph AC_K.

1-3.1 MEASURING POWER AND ENERGY

Power Versus Energy

Power is the rate of flow of energy. This is true for any form of energy, not just electricity. If you wish to boil a cup of water you need a quantity of **energy** to get the job done, about 30 watt-hours. Any specific power level, say a thousand watts (kilowatt, or kW), may or may not make you a cup of tea depending on how long the power continues to flow. A typical microwave oven delivers power at a rate of about 1 kW (not 1 kW per hour). If it heats your water for one second, the water will receive *power* at the *rate* of one thousand watts, but it will gain very little *energy* and it will not make tea. Two minutes in the micro-wave will deliver the necessary energy, $\frac{1}{30}$ of a kWh.

Unit Arithmetic

Units—kilowatts, hours, and dollars—follow the normal laws of arithmetic. But it must be understood that a kWh means a (kW × h) and a $ per hour means a ($/h).

Also note that "8760 hours per year" has the value of 1, because it equals (8760 h)/(1 year), and (8760 h) = (1 year).

As an example, \$100/kWy =

$$\frac{\$100}{\text{kW}\times\text{year}}\times\frac{1000\text{ kW}}{1\text{ MW}}\times\frac{1\text{ year}}{8760\text{ h}}$$

which reduces to \$11.42/MWh.

Confusion arises because it is more common to have the time unit in the measurement of a flow than in the measurement of a quantity. Thus if you want to fill your gas tank, you buy a quantity of 15 gallons of gasoline, and that flows into your tank at the rate of 5 gallons per minute. But if you need a quantity of electric energy, that would be 30 watt-hours, and it would be delivered at the rate of 1000 watts.[2] Because a watt-hour is a unit of energy, it would make sense to speak of delivering 1000 watt-hours per hour, but that just boils down to a rate of 1000 watts (1 kW) because a watt-hour per hour means watts times hours divided by hours, and the hours cancel out.

The Price of Power and Energy

Because power is a flow, its total cost is measured in dollars per hour, not dollars. The total cost of a certain quantity of energy is measured in dollars. Consequently the *price* (per unit cost) of power is measured in dollars per hour per MW of power flow, while the price of energy is measured in dollars per MWh. But these units are the same:

$$\text{(dollars per hour) per MW} = (\$/\text{h})/\text{MW} = \$/\text{MWh}$$

so the units for the price of power are the same as for the price of energy.

Typically the price of retail energy is about 8¢/kWh.[3] At that price, the price of power would be 8 cents/hour for a kilowatt of power flow, which is the same. These units are convenient for home use but are inconveniently small for bulk power systems. Consequently this book will use megawatts (millions of watts) instead

2. Watts per hour has units of watts divided by hours and has no use in the present context.

3. Average revenue per kilowatt hour to ultimate residential consumers was 8.06¢/kWh, according to DOE (2001c, Table 53).

of kilowatts. The same energy price can be re-expressed as $80/MWh. When discussing large markets and annual energy use, power may be measured in gigawatts (GW, or billions of watts) and energy in terawatt hours (TWh, or trillions of watt hours).

Another commonly used unit is the **mill**, short for "milli-dollar," or $\frac{1}{1000}$ of a dollar. This unit might seem particularly inappropriate for wholesale markets, but it is commonly used to compensate for using the kW which is also inappropriately small. Together these give rise to "milli-dollars per kilowatt-hour," often incorrectly shortened to "mills." Scaling both the numerator and denominator up by 1000 has no effect on the numeric value and converts milli-dollars to dollars and kilowatts to megawatts. So 80 mills/kWh is identical to $80/MWh.

1-3.2 MEASURING GENERATION CAPACITY

The size of a generator is measured by the maximum flow of power it can produce and therefore is measured in MW. The capacity to produce a flow of power is best conceptualized as a flow just as a MW of power is a flow of energy.[4]

In principle one could define an amount of capacity related to the flow of capacity as energy is related to power, but this is not necessary. Moreover, it is likely to cause confusion because when applied to a generator, it would aggregate a flow of capacity over many years without any discounting. For these reasons, the idea of a capacity amount, different from a capacity flow, will not be introduced or utilized.

Having found that capacity, like power, is a flow measured in MW, it is natural to ask if it is priced in $/MWh as is power. Most would say no, but it is best to look to its use in solving real economic problems before drawing this conclusion. Consider the problem of choosing which generator can most cheaply serve a load of a particular duration. The long tradition of solving this problem by using "screening curves" will provide the key to this puzzle.

1-3.3 PRICING GENERATION CAPACITY

The "Overnight" Cost of Capacity

A generator has an "overnight cost" which is typically given in $/kW. For example, the overnight cost of a coal plant might be $1,050/kW, so a 1000 MW plant would cost $1,050 million. In economic terms, this is the present-value cost of the plant; it would have to be paid as a lump sum up front to pay completely for its construction.

A conventional gas-turbine generator (GT) would have an overnight cost closer to $350/kW. Although the GT is three times cheaper than the coal plant, for some

4. The flow of available capacity is interrupted during generator outages, but the flow of installed capacity is continuous. This chapter ignores the difference and assumes that the flow of capacity from a generator is continuous and constant.

purposes the coal plant is the more inexpensive choice. Fuel costs must always be taken into account when evaluating the choice of generators. Coal plants are built because their cost of fuel per unit of energy output is less. Assume coal costs only $10/MWh of energy produced, while the cost of fuel for a GT comes to $35/MWh. Now which plant is cheaper?

More information is needed. The comparison depends on how much the plant will be used, and that depends on the load it will serve. For concreteness, assume that the load has a duration of 25% (2190 hours/year) so the plant serving it will have a capacity factor of 25%. Now, which plant is cheaper?

Focusing on only the basics, the problem seems workable. The overnight cost captures the *fixed cost* of generation, and the fuel cost per unit of output captures the *variable cost*. Duration gives a sufficient description of the load. But the problem is still impossible to solve because the fixed cost of capacity has been measured in the wrong units. Overnight costs measured in $/kW cannot be added to fuel costs measured in $/kWh. This would produce nonsense.

Fallacy **1-3.1** **Fixed and Variable Costs Are Measured in Different Units**

Because capacity is usually paid all at once, while fuel is paid for over time, variable costs but not fixed costs should include a time dimension.

When units have the same "dimensions," they differ only by a *scale factor* (a pure number). Different quantities having units of the same dimension can be added. For example, 1 MWh can be added to 100 kWh to get 1100 kWh (or 1.1 MWh). But quantities whose units have different dimensions cannot be added. This is the meaning of the famous saying, "you can't add apples and oranges." For example, 1 MW cannot be added to 1 MWh. Engineers and physicists pay close attention to mismatched units because they always signal deeper trouble. Any calculation that involves adding MW and MWh simply does not make sense.

Identifying Fixed Costs on Screening Curves

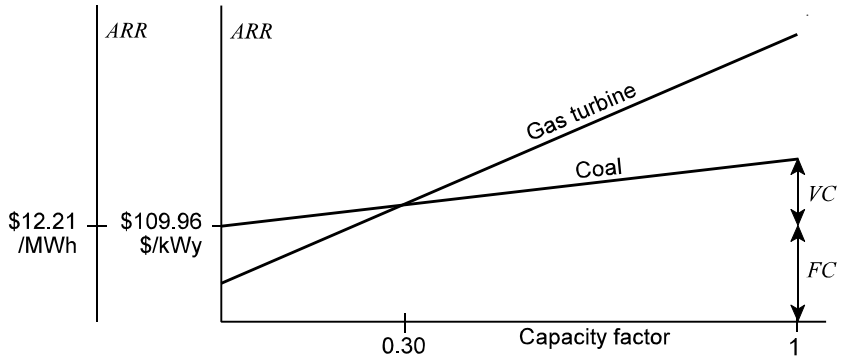
Screening curves, shown in Figure 1-3.1, are used to compare generation costs by taking account of the three factors of our present problem: fixed cost, variable costs and load duration (which determines the generator's capacity factor). Necessarily, they provide guidance on the proper units for fixed costs. Traditionally, these curves plot "**annual revenue requirement** per kW" (*ARR*) as a function of capacity factor (*cf*). The generator's **capacity factor** is its percentage utilization which is determined by the load's duration.[5]

Traditionally, the **variable cost** component of *ARR* is computed by taking the fuel cost expressed in $/MWh and converting to $/kWy.[6] The result is $87.60/kWy for a coal plant and $306.60/kWy for a GT. This assumes full-time operation, so

5. Duration is measured as a percentage (see Chapter 1-4), so if all load served has the same duration, the capacity factor equals load duration. If the load has a range of durations, these must be averaged.

6. For simplicity, this assumes that fuel is the only variable cost. Operation and maintenance include an additional variable cost component which should be expressed in $/kWy.

Figure 1-3.1
Use of screening curves to select a generator.



to find the variable component for any particular *cf*, these must be multiplied by *cf*, 25% in the case of the present example.

The overnight cost of capacity is more problematic. A coal plant with an overnight cost of $1,000/kW does not cost $1,000/kWy. This would imply a plant life-time of one year and a discount rate of zero. The correct fixed-cost component of *ARR* is the overnight cost amortized ("levelized") over the life of the plant. This is equivalent to computing home mortgage payments based on a mortgage that lasts the life of the house. Obviously a discount rate (interest rate) is involved. The formula for amortization is

$$FC = \frac{r \cdot OC}{1 - e^{-rT}} \approx \frac{r \cdot OC}{1 - 1/(1+r)^{T}} \tag{1-3.1}$$

Notice that **fixed cost** (*FC*) depends only on overnight cost (*OC*), the discount rate (*r*, in % per year) and the life of the plant (*T*, in years).[7]

Table 1-3.1 Technology Costs

Technology	*VC* (/MWh)	*VC* (/kWy)	*OC* (/kW)	*FC* (/kWy)	*FC* (/MWh)
Gas turbine	$35	$306.60	$350	$40.48	$4.62
Coal	$10	$87.60	$1050	$106.96	$12.21

Fixed-costs are based on $r = 0.1$ and on $T = 20$ for gas turbines and 40 for coal plants. Equation 1-3.1 gives fixed costs in $/kWh which are then converted to $/MWh by dividing by 8.76.

FC is a constant flow of cost that when added to *VC* gives *ARR*, the annual revenue requirement per kW of generation capacity. Of course this assumes a capacity factor of 1. If *cf* is less, *VC* will be reduced proportionally, but *FC* is unaffected because capacity must be paid for whether used or not. That is why *FC* is termed the *fixed* cost. The formula for *ARR* is

$$\text{Screening Curve: } ARR = FC + cf \times VC$$

7. Using *monthly* instead of *annual* compounding in the second formula greatly improves its accuracy as an approximation. To do this, change *r* to $r/12$ in the denominator and *T* to 12 *T*.

for *ARR* to be valued in \$/kWy, both *FC* and $cf \times VC$ must also be valued in \$/kWy. As these are the traditional units for *ARR*, the traditional units for fixed cost must also be \$/kWy. These units have the same dimension as \$/MWh and any quantity expressed in \$/kWy can be converted to \$/MWh by dividing by 8.76.

Variable cost is naturally expressed in \$/MWh, so capacity factor, *cf*, must be a pure number (dimensionless), otherwise, $cf \times VC$ would not have the same units as *ARR*. This is correct; a capacity factor is just the fraction of a generator's potential output that is actually produced. It is actual energy output divided by potential energy output, so the energy units cancel.

Result 1-3.2 **Energy, Power, and Capacity Are Priced in \$/MWh**
Although power is measured in MW and energy in MWh, both are priced in \$/MWh. Like power, generating capacity is a flow measured in MW and consequently is also priced in \$/MWh.

The Rental Cost of Capacity

Fixed costs are the costs of generation capacity. It may be argued that buying a generator is buying capacity and that generators are measured in MW, not in MWh. This is only partially true. If a 1 MW gas-turbine generator is worth \$350,000, does this mean 1 MW of capacity is worth \$350,000? No, the gas turbine is worth that only because it has a certain expected lifetime. An identical but older gas turbine is worth less, even though it has the same 1 MW capacity. Thus the price of capacity always involves a time dimension, either explicitly or implicitly.

Measuring capacity in MW indicates that capacity is being considered a flow. A 100-MW generator delivers a 100-MW flow of capacity for some unspecified period of time. That flow must be paid for by a flow of money—so many dollars per hour. This corresponds to a *rental cost*. If a generator is rented, the cost of renting will be so much per hour, or per day, or per year. If this is scaled by the generator's capacity, for easy comparison with the rental rate of other generators, then it is natural to express the rental cost of a generator in \$/h per MW, or equivalently in \$/MWh.

The above screening curve analysis can be summarized as saying that generation capacity costs should be expressed as a rental rate and not as a one-time (overnight) purchase price. Rental rates naturally have the same units as variable costs and so make total and average cost calculations convenient.

Two Kinds of Average Cost: Avoiding Confusion

The cost of operating a generator with a specific capacity factor can be read from a screening curve. Traditionally this cost is expressed in \$/kWy and called an annual cost. Although a kWy has the dimensions of energy, this cost is **not** the annual cost of *energy* produced by the plant! A screening curve shows the average cost of using the coal plant's *capacity*.

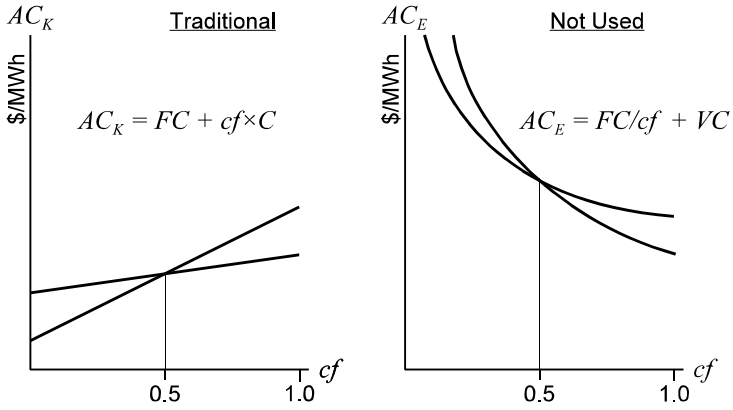


Figure 1-3.2
Capacity-cost based and energy-cost based screening curves.

With the price of energy always expressed as an hourly cost, it is more convenient to divide the annual cost, in $/kWy, by 8.76 and arrive at an average cost per hour for the year expressed in $/MWh. Like the annual cost, this hourly average cost is still not a cost of energy produced but the cost of using capacity. This book will always report capacity, energy and power costs in $/MWh for ease of comparison and to make addition of costs and cost averaging possible.

Although screening curves plot the average cost of capacity use, the average cost of energy produced is also interesting and could be used to construct hyperbolic screening curves. A pair of these is shown on the right side of Figure 1-3.2; they are nonlinear (hyperbolic), but they still intersect at exactly the capacity factor at which one plant becomes more economical than the other. The equations for the linear and hyperbolic screening curves are closely related and are shown in Figure 1-3.2.

The average cost of capacity (AC_K) used with capacity factor *cf* is the fixed cost of using that capacity, plus *cf* times the variable cost of producing energy. If $cf = \frac{1}{3}$, then one third of the variable cost of maximum potential energy output must be added to the constant fixed cost which increases the average cost per unit of capacity by $cf \times VC$.

Screening Curves

A traditional screening curve plots average cost as a function of capacity factor. When using the screening curve equation, this average cost can be confused with the average cost of the *energy* produced by a certain type of generator. Instead, it is the average cost of using a unit of capacity.

Screening curves could have been defined using the average cost of energy. Then as the capacity factor approached zero the average cost would approach infinity. The average-energy-cost equation is used to analyze market equilibria in later chapters, but nonstandard screening curves are never used.

The average cost of energy (AC_E) when the generator runs with capacity factor *cf* is the variable cost of producing that energy, plus the fixed cost of the capacity divided by *cf*. If $cf = \frac{1}{3}$, then fixed costs must be spread over only $\frac{1}{3}$ of the total possible energy output, so they are multiplied by 3 (divided by *cf*) before being added to *VC*.

A load slice is a horizontal strip cut from a load-duration curve (see Chapter 1-4.1). Depending on its average duration it will be served by some particular technology, baseload, midload, or peaking. Any given load slice is defined by a capacity, K_{slice}, which is the height of the slice, and an average duration, *D*. It also has a total energy requirement, $E = D \times K_{slice}$. To serve

this load, generation capacity of K_{slice} must be installed and must run with a capacity factor of $cf = D$. Having selected a technology, one can compute the **average cost** per MW-of-*capacity* of serving the load, AC_K, and the average cost per MWh-of-*energy* of serving the load, AC_E. The total cost of serving load is then given by both $K_{slice} \times AC_K$ and $E \times AC_E$. Because K_{slice} and E are both fixed, choosing the technology that minimizes either AC_K or AC_E will minimize the total cost of energy. This is why either the traditional or the hyperbolic screening curves can be used.

These relationships can be summarized as follows. For a particular load slice served by generators with fixed cost, FC, and variable cost, VC, the average cost of capacity and energy can be found as follows:

$$\text{Capacity: } AC_K = FC + cf \times VC = FC + D \times VC \tag{1-3.1}$$

$$\text{Energy: } AC_E = FC/cf + VC = FC/D + VC \tag{1-3.2}$$

The capacity factor of the generator, cf, equals the average duration of the load, D.

No one uses hyperbolic screening curves, but when an average cost is computed for a specific technology, say by DOE, a value for AC_E (not AC_K) is always computed. Typically, DOE might report the overnight cost (OC), some information about fuel costs, and a value for AC_E based on technical capacity factor (cf).[8] In other words, DOE reports some technologically determined value on the technology's *hyperbolic* screening curve.

AC_K is used to determine the optimal durations of various generation technologies, and from these durations the optimal investment in these technologies. Since competitive markets optimize technology, AC_K is also used to determine competitive outcomes. Either AC_E or AC_K may be used to compare the cost of peak energy with the value of lost load, depending on whether **peaker** costs are equated to the value of lost load or the average hourly cost of lost load (see Chapter 2-2 and 2-3). AC_E is also well suited to DOE's interest in alternative technologies—nuclear, wind, solar, and so on. These have in common capacity factors which, even in a market environment, are not affected by normal variations in market structure but are instead technologically determined because their variable costs are almost always below the market price. They run whenever they are physically able, so their capacity factor is determined by their technical capability. The economics of an alternative technology can be assessed by comparing its AC_E with the market's average price.

Standard technology generators have capacity factors determined by the market and not just by their technology. In this case, the duration of the load they serve, which determines their capacity factor, needs to be determined from their fixed and variable costs along with those of other technologies. This is done with screening curve analysis, or with algebra based on screening curves. Traditional linear screening curves prove simplest. These curves and their associated algebra will be used throughout the book, as is the formula for AC_E.

---

8. See DOE (1998a) Tables 14–17. In a table of cost characteristics of new generating technologies (DOE 2001a, Table 43) *OC* is given, but not *FC*. In a slide labeled "Electricity Generation Costs," DOE (2001b) reports the capital costs in mills/kWh. As the title indicates, these are FC/cf, for cf determined by technology-based capacity factors, and so are components of AC_K, as advertised in the title and thus points on the hyperbolic screening function.

Table 1-3.2 Fixed and Variable Cost of Generation

Type of Generator	Overnight Capacity Cost $/kW	Fixed Cost $/MWh	Fuel Cost $/MBtu	Heat rate Btu/kWh	Variable Cost $/MWh
Advanced nuclear	1729	**23.88**	0.40	10,400	**4.16**
Coal	1021	**14.10**	1.25	9,419	**11.77**
Wind	919	**13.85**	—	—	**0**
Advanced combined cycle	533	**7.36**	3.00	6,927	**20.78**
Combustion turbine	315	**4.75**	3.00	11,467	**34.40**

* Overnight capacity cost and heat rates are from DOE (2001a), Table 43. Plants not labeled "advanced" are "conventional." Rental capacity costs are computed from overnight costs, a discount rate of 12% and assumed plant lifetimes of 40 years except for wind and gas turbines which are assumed to be 20 years. For simplicity, operation and maintenance costs are ignored.

1-3.4 TECHNICAL SUPPLEMENT

Checking Fixed-Cost Units with the Amortization Formula

As a final check on the units of fixed costs, the amortization formula can be analyzed. Interest rates (e.g., 10% per year) has the dimension of "per unit time," and T has the dimension of time, so rT is dimensionless, that is, a pure number. This is necessary for compatibility with "1" in the denominator. In the numerator, OC has the traditional units of $/kW and "r" again has the dimension of 1/time, so $r \times OC$ has the dimensions of OC per unit time, for example, $/kW per year. If overnight cost is measured in $/kW and interest is given in percent per year, fixed cost must be measured in $/kWy.

Fixed and Variable Costs for Different Technologies

Table 1-3.1 computes fixed and variable costs for five types of generators as an example of converting overnight cost to fixed costs. The listed values of FC and VC are exactly the values needed to draw screening curves and choose the most efficient plant to serve loads of any duration. For example, the cost of serving load of duration D with an advanced combined-cycle plant is

$$AC_K = (7.36 + 20.78\,D)\ \$/\text{MWh}.$$

To convert this to the more traditional units of $/kWy, both values should be multiplied by (1 M/1000 k)(8760 h/y) or 8.76. (Note that, including units, this is just multiplication by 1 since 1 M = 1000 k and 8760 h = 1 y.)

To avoid having AC_K appear to have the same units as variable costs, its units are often stated as "$/kW per year" which translates to $/kW/year. But just as $x/y/z = x/(y \times z)$, so $/kW/year equals $/kW-year which is denoted by $/kWy. The phrase "$/kW per year" is correct, but it means no more and no less than $/kWy, which has the dimensions of dollars per energy.

Chapter 1-4

Power Supply and Demand

And when the Rain has wet the Kite and Twine, so that it can conduct the Electric Fire freely, you will find it stream out plentifully from the Key on the Approach of your Knuckle.

Benjamin Franklin
1752

THE PHYSICAL ASPECTS OF SUPPLY AND DEMAND PLAY A PROMINENT ROLE IN POWER MARKETS. Shifts in demand, not associated with price, play a role in all markets, but in power markets they often receive attention to the exclusion of price. This is not simply the result of regulatory pricing; even with market prices, demand shifts will play a key role in determining the mix of production technologies. In this way hourly demand fluctuations determine key long-run characteristics of supply.

Because electric power cannot be stored, production always equals consumption, so the difference between supply and demand cannot be indicated by flows of power. Neither is the instantaneous difference indicated by contracts since real-time demand is determined by customers physically taking power. The short-run supply–demand balance is indicated by voltage and, especially, frequency. This unusual market structure requires some elementary background in system physics. More detail is provided in Chapters 5-1 and 5-2.

Chapter Summary 1-4: Load duration curves are still relevant in unregulated markets, but their role in analysis is more subtle because their shape is affected by price and its correlation with load. They can still be used with screening curves to check an equilibrium, but to predict an equilibrium they must be used in combination with price elasticity.

Power production always equals consumption (counting losses as part of consumption) which makes it impossible to assess the supply–demand balance by observing quantities or quantity flows. Instead, frequency is the proper indicator of system-wide balance, and net unscheduled flows between regions are used to share the responsibility of maintaining this balance.

Section 1: Describing the Demand for Power. A year's worth of hourly fluctuations can be usefully summarized by a load-duration curve that plots demand against duration, the fraction of the year during which demand is at or above a certain level. It can also be thought of as the probability of finding load above a certain level.

If customers are charged real-time prices, peak demand will be reduced, allowing a reduction in generating capacity. The result will be a load-duration curve with its peak cut off horizontally.

Section 2: Screening Curves and Long-Run Equilibrium. If the screening curves of the available technologies are drawn on the same graph, their intersections determine capacity factors that mark technology boundaries. By mapping these capacity factors to durations and then to the load-duration curve, the optimal capacities for these technologies can be read off the vertical axis.

This technique can be used to partially confirm a market equilibrium but not to find one. In a market, price affects the shape of the load-duration curve, so it cannot be taken as given until the equilibrium is known.

Section 3: Frequency, Voltage, and Clearing the Market. When consumers turn on ten 100-W light bulbs, they are demanding 1000 W of power, and if generators supply only 900 W, the system will not be "in balance." In spite of this, the power supplied will exactly equal the power consumed (ignoring losses). This equality of power flows is caused by a decrease in voltage sufficient to cause the 100 W bulbs to use only 90 W of power. For motors the same effect is caused by a drop in frequency. Because voltage is automatically adjusted at substations, frequency is the main balancer of power inflows and outflows.

The United States is divided into three AC interconnections: the Western, Eastern, and Texas. The system frequency is constant throughout each AC interconnection, which means that a change in frequency cannot be used to locate a supply–demand imbalance. Instead, net power flows are tracked out of each control area and compared with scheduled power flows. This allows the imbalance to be located.

1-4.1 DESCRIBING THE DEMAND FOR POWER

Traditionally the demand for power has been described by a **load-duration curve** that measures the number of hours per year the total **load** is at or above any given level of demand. An example is shown in Figure 1-4.1. Total demand (load) is a demand for a flow of power and is measured in MW. Although the load-duration curve describes completely the total time spent at each load level, it does not include information about the sequence of these levels. The same load-duration curve can

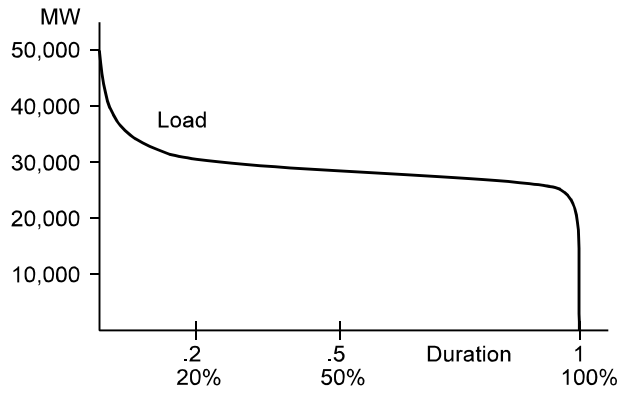


Figure 1-4.1
A load-duration curve.

be produced by wide daily swings in demand and little seasonal variation or by wide seasonal variation and limited daily swings.

The introduction of a market adds the dimension of *price*. Economists often represent demand by a *demand curve* which expresses demand solely as a function of price. Nonprice fluctuations of the type captured by a load-duration curve are referred to as shifts in the demand curve and are generally not described in detail. But electricity is different because it is not storable, so peak demand must be satisfied by production from generators that are used as little as 1% of the time. Such generators, **peakers**, are built with technology that differs markedly from that used for **baseload generators** which run most of the time and are stopped only rarely. As a result, power markets face the problem of determining how much generation capacity should be built using each type of technology, for example, coal-fire steam turbines or gas-fired combustion turbines (**gas turbines**). This explains the unusual importance of demand shifts and consequently of load-duration curves in power markets.[1]

Load-Duration Curves

A load-duration curve can be constructed for a given region (or for any collection of loads) by measuring the total load at hourly intervals for each of the 8760 hours in a year, sorting them, and graphing them starting with the highest load. The result is a curve that slopes downward from the maximum load in the peak hour, hour 1, to the minimum load, **baseload**, in the most off-peak hour, hour 8760 (see Figure 1-4.1).

Duration is traditionally measured in hours per year, but both hours and years are measures of time, so duration is **dimensionless**, which means it can be expressed as a pure number, a ratio, or percentage. To convert from units of hours per year (h/year) to a pure number, simply multiply by 1 in the form (1 year)/(8760 h). Duration has a natural interpretation as the probability that load will be at or above a certain level. To use this interpretation pick a load level, say 35 GW, and using the load-duration curve, find the corresponding duration, 20% in this case. This

1. Service industries such as restaurants and airlines often have demand fluctuations which cause similar problem because their output is not storable, but they tend to use the same technology on and off peak.

indicates that load is 35 GW or greater 20% of the time. Put another way, the probability of load being 35 GW or greater in a randomly selected hour is 20%. This interpretation is most convenient.

The Price-Elasticity of Demand

Presently, demand is almost completely unresponsive to price in most power markets because wholesale price fluctuations are not usually passed on to retail customers. Often retail prices remain under some form of price regulation, but competitive retailers have also been slow to implement real-time pricing. In the longer run, retail prices do change, sometimes seasonally. In the long run a 10% increase in the price of power will cause approximately a 10% reduction in the use of power.[2] This is not a very accurate approximation, but the long-run response to a 10% increase in price is likely to be found between 5% and 15% and is certainly not zero. Economists term this price sensitivity a **price elasticity of demand**, which is often shortened to *demand elasticity*. If a 10% change in price causes a 5%, 10%, or 15% change in demand, the elasticity is said to be 0.5, 1.0, or 1.5, respectively. (Technically, demand elasticities are negative, but this book will follow the common convention of re-defining them to be positive.)

Real-Time Pricing and the Load-Duration Curve

Under regulation, residential load usually faces a price that fluctuates very little while commercial and industrial load often face time-of-use (TOU) pricing or demand charges. Time-of-use prices are designed to be high when demand is high, but the approximation is crude as they are set years in advance. Consequently they miss the crucial weather-driven demand fluctuations that cause most problematic supply shortages. Demand charges are no more accurate as they are based on individual demand peaks, not system peaks. Coincident-peak charges improve on this by charging customers for their use at the time of the system peak, but these are less common.

Because supply is fairly constant, the market is tightest when demand is highest. Consequently, high wholesale prices correspond well with high demand. If these real-time prices are passed through to customers, then retail prices will track load fairly well. Although real-time prices work best, all four pricing techniques, TOU, demand, coincident-peak, and real-time, tend to raise prices when demand is highest and reduce prices when demand is lowest. This results in lowering the peak of the load-duration curve and raising the low end of the curve.

If load faced real-time prices, the need for generation capacity might be reduced to the point where the load-duration curve under regulation had a duration of, say, 10%. Then, between 0% and 10% duration, supply and demand would be balanced by price. Instead of having generation follow load, load would be held constant by price at the level of installed capacity. In the lowest duration hours, price would

2. There is no natural definition of short- and long-run demand elasticity, which can be defined usefully over any time horizon from five minutes to twenty years. This text will use short-run elasticity to mean something on the order of one day and long-run to mean about five years.

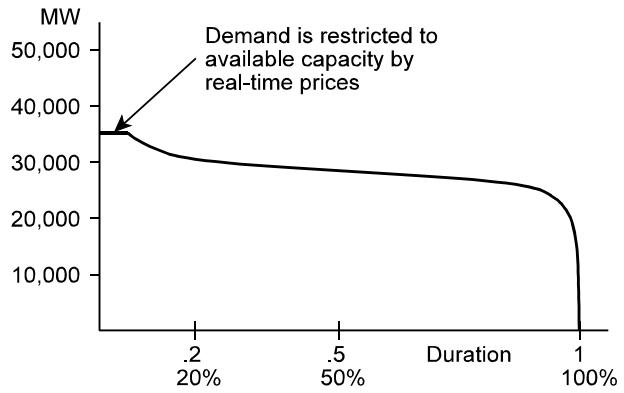


Figure 1-4.2
The effect of price elasticity on load duration.

need to be very high to reduce demand to this level. By fluctuating sufficiently, price would control demand and produce a flat-topped load-duration curve with a maximum load just equal to generating capacity as shown in Figure 1-4.2.

1-4.2 SCREENING CURVES AND LONG-RUN EQUILIBRIUM

When demand is inelastic or when it faces a fixed price so that the load-duration curve is fixed, this curve can be used to find the optimal mix of generation technologies. The technique was developed for a regulated power system in which price and the load-duration curve are often fixed, but it is still useful for understanding certain aspects of competitive markets.

It assumes that fixed and variable costs adequately describe generators. These are used to draw screening curves for each technology on a single graph as shown in Figure 1-4.3. The intersections of these curves determine capacity factors that separate the regions in which the different technologies are optimal. These capacity factors equal the load durations that determine the boundaries between load that is served by one technology and the next. The screening curves in the figure are taken from Figure 1-3.1 and intersect at a capacity factor, *cf*, of approximately 30%. Consequently all load with a duration greater than 30%, or about 2600 hours, should be served by coal plants, while load of lesser duration should be served by gas turbines. The arrow in the figure shows how the needed capacity of coal plants can be read from the load-duration curve. The optimal GT capacity is found by subtracting this from maximum load which is the total necessary capacity. (Forced outages and operating reserve margins are considered in Parts 2 and 3.)

If customers face the wholesale market price through real-time pricing, this technique cannot be used because the load-duration curve depends on price, and price depends on the choice of technology, and the choice of technology depends, as just described, on the load-duration curve. This circularity is in no way contradictory, but it makes it difficult to find the competitive market equilibrium. Not only is calculation more difficult, but, also, the elasticity of demand must be known.

In spite of this circularity, the traditional technique can be used to partially confirm a long-run equilibrium. The load-duration curve observed in a market

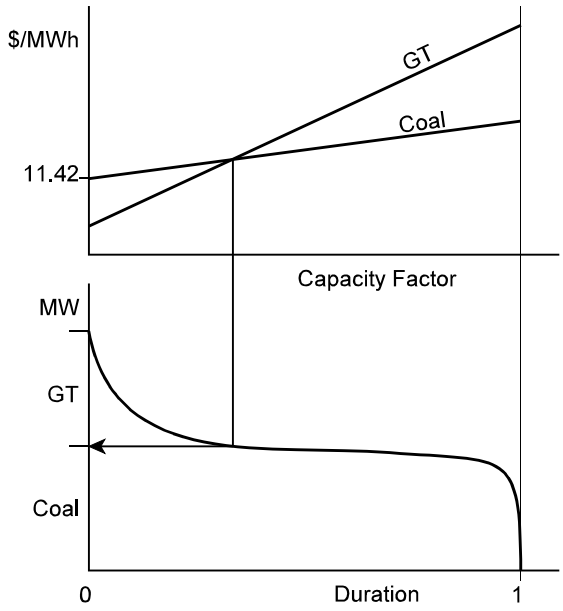


Figure 1-4.3
Using screening curves to find the optimal mix of technologies.

includes the effect of price on demand. When it is used along with the screening curves of the available technologies, the traditional method should predict the mix of technologies observed in the market if the market is in long-run equilibrium. In practice many complications must be overcome.

1-4.3 FREQUENCY, VOLTAGE, AND CLEARING THE MARKET

So far, this chapter has considered how to describe demand and how to find the optimal mix of technology to supply it. This section considers the physical details of the supply–demand balance in real time. At any instant, customers are using power, generators are producing it, and the amount produced is exactly equal to the amount consumed. Some may object to the word "exactly," but the discrepancy is at least a thousand times smaller than anyone's ability to measure it and is entirely irrelevant. The determination of the supply–demand balance depends on electrical phenomena more subtle than the concepts of quantity and quantity flow.

Losses

In real networks, a few percent of the power consumed is consumed by the network. This consumption should be considered part of demand even though it serves no end used. With this convention, the system can be viewed as maintaining a perfect balance between supply and consumption (including losses) at all times and between supply and demand whenever customers are getting the power they want.

Convention	**Loss Provision is Not Considered Part of Supply or Demand** Losses will be considered as a service paid for by traders and provided separately from the trading arrangement. Consequently, from a trading point of view, power flows can be viewed as lossless.

Supply precisely equals consumption because there is no storage of power in transmission systems.[3] But if supply equals consumption regardless of price, what signal should be used for price adjustment? How can demand be observed to be either greater than or less than supply? A mismatch between supply and demand is signaled not by power flows but by frequency and voltage. When they are below their target values, demand exceeds supply and vice versa.

Frequency and Voltage

Power systems attempt to maintain a constant **frequency**, the rate at which alternating current alternates. In the United States, alternating current (AC) reverses direction twice, thus completing one *cycle* and returning to its original direction, 60 times per second. The frequency of AC in the United States is therefore said to be 60 cycles per second, also known as 60 Hertz or 60 Hz. In many countries the target frequency is 50 Hz.

Voltage is the amount of electrical pressure that pushes current through electrical appliances such as lights and motors. As with frequency, power systems have a certain target voltage that they attempt to maintain. In the United States the target residential voltage is about 120 volts. In some countries the target voltage is about twice as high. When an unprotected 120-V appliance is plugged into a 240-V outlet, the extra electrical pressure (voltage) causes twice as much current to flow through the appliance. This causes the appliance to use four times as much power (power is voltage times current) and the appliance typically burns out. The important point for this section is that as voltage increases, most appliances use more power, and as voltage decreases most appliances use less power.

Imagine a system with ten generators operating at full throttle supplying ten thousand homes with lights burning and motors running. If one generator goes off line, two things happen. The system voltage and frequence both decrease. Both cause electrical appliances to use less power. This effect has been described for voltage, but for more complex reasons most motors use less power when the system frequency declines. The decline in voltage and frequency is produced automatically by the physics of the entire system including all loads and generators. It happens to the exact extent necessary to balance inflow (supply) and outflow (consumption). If this did not happen, a law of physics, just as fundamental as the law of gravity, would be violated.

Although nothing can prevent the combined drop in frequency and voltage when generation is reduced and load maintained, it is possible to influence the relative extent to which each decreases. In fact, the system has automatic controls that do

3. More precisely, the amount stored is minuscule and cannot be utilized for trade.

just that. At substations, where very high transmission voltages are reduced to the lower, but still high, distribution voltage, there are automatically controlled transformers. These adjust so that the distribution voltage remains relatively constant even when the transmission voltage drops. Because of these devices, more of the power flow adjustment that accompanies the loss of a generator is accomplished through frequency reduction than through voltage reduction. Nonetheless both can and do happen.

Frequency and Interconnections

An interconnection is a portion of the power grid connected by AC power lines. The three interconnections in the United States—the Eastern **Interconnection**, the Western Interconnection, and most of the Great State of Texas—each maintain a uniform frequency. Frequencies in Maine and Texas bear no particular relationship to each other, but the AC voltage in Maine stays right in step with the AC voltage in Florida, night and day, year in and year out. The frequency in every utility in an entire interconnection is exactly the same.[4] If one utility has a problem, they all have a problem.

These three interconnections are connected to each other by a number of small lines, but they are separate interconnections because the connecting lines are all DC lines. No AC power flows between them. On DC lines, the electrical current flows in only one direction; it does not alternate directions. Thus DC lines have no frequency and as a consequence need not (and cannot) be synchronized with the power flow of an AC interconnection. This allows trade between two different interconnections that are not synchronized with each other.

The Signal for Price Adjustment

When a generator breaks down unexpectedly (a forced outage) and supply decreases, demand is then greater than supply, even though consumption still precisely equals supply. Consumption is less than demand because of rationing. A consumer with a 100-W light turned on is demanding 100 watts of power. During a brown out, however, 100 watts are not supplied to the bulb as power to the bulb is rationed by the suppliers. This rationing is not due to deliberate action but is a consequence of system physics which automatically lowers the voltage and frequency. For simplicity, in the remainder of this section, rationing will be discussed as if it happened solely through frequency reduction, as this is generally considered to be the predominant effect.

A drop in frequency below the target level of 60 Hz is a clear and accurate indication that demand exceeds supply for the interconnection as a whole. Similarly, any frequency above 60 Hz indicates that supply exceeds demand. In other words, more than 100 watts are being delivered to 100-W motors. This extra power is

4. If the frequency difference between Maine and Florida were 0.001 Hz, for one minute, it would cause an accumulated phase change between the two states of 22 degrees. This would lead to dramatic changes in power flow. Thus while the frequency lock between utilities is not exact, it is extremely tight, and there can be no persistent frequency difference. One utility cannot have a problem unless they all do.

generally unwanted because appliances are built and operated on the assumption that power will be delivered at a frequency of 60 Hz.

Because frequency indicates the discrepancies between supply and demand, frequency is the right guide for interconnection-wide price adjustment. When frequency is high, price should be reduced; when frequency is low, price should be raised. This is the classic adjustment process for keeping supply equal to demand.

Definition **Demand**

The demand for power is the amount of power that would be consumed if system frequency and voltage were equal to their target values for all consumers. Note that shed load is included as part of demand. This is an economic definition and contradicts the engineering definition provided by North American Electric Reliability Council (NERC). (Often "load" is used to mean demand.)

Result 1-4.1 **Supply Equals Consumption but May Not Equal Demand**

As in all markets, demand is the amount customers would buy at the market price were supply available. If voltage or frequency is low, customers consume less power than they would like so supply is less than demand.

As always, the real world adds one more layer of complexity. The frequency in every power market in an interconnection is exactly the same. Thus frequency reveals nothing about the supply and demand conditions in any particular market but only about the aggregate supply and demand conditions of the entire interconnection. Consequently individual markets cannot rely on the frequency alone to determine their price adjustments.

NERC defines another control variable that takes account of both frequency and the net excess flow out of a trading region (the net interchange). The net excess outflow is the actual outflow minus the scheduled outflow. An excess outflow is a strong signal that supply is greater than demand in the trading region. If the frequency is high in the interconnection this is a weak signal of excess supply in any particular market. These two signals are combined to form a single indicator of excess supply for each market. The indicator is called the area control error, or **ACE**. Control areas are required to keep their ACE near zero, and they do. ACE is the main indicator of the supply–demand balance in every control area in the United States and when there is a market, it is the signal that determines whether the price will be increased or decreased by the system operator.

Chapter 1-5

What Is Competition?

The rich, . . . in spite of their natural selfishness and rapacity, . . . though the sole end which they propose . . . be the gratification of their own vain and insatiable desires, they divide with the poor the produce of all their improvements. They are led by an invisible hand to make nearly the same distribution of the necessaries of life, which would have been made, had the earth been divided into equal portions among all its inhabitants, and thus without intending it, without knowing it, advance the interest of the society.

Adam Smith
The Theory of Moral Sentiments
1759

COMPETITION IS LEAST POPULAR WITH THE COMPETITORS. Every supplier wants to raise the market price, just as every buyer wants to lower it. Perfect competition frustrates both intentions.

Some commodity markets provide almost perfect competition; eventually power markets may work almost as well. But designing such markets is difficult. Economic competition is not like competition in sports, which may be considered perfect when there are just two powerful and equal competitors. Economists consider competition to be **perfect** when every competitor is small enough (*atomistic* is the term used) to have no discernable influence against the "invisible hand" of the market.

Adam Smith guessed intuitively that a perfectly competitive market, in the economic sense, would produce an outcome that is in some way ideal. Many difficulties can cause a market to fall short of this ideal, but even a market that is only "workably competitive" can provide a powerful force for efficiency and innovation.

Power markets should be designed to be as competitive as possible but that requires an understanding of how competition works and what interferes with it. On its surface, competition is a simple process driven, as Adam Smith noted, by selfishness and rapacity; but the invisible hand works in subtle ways that are often misunderstood. Those unfamiliar with these subtleties often conclude that suppliers are either going broke or making a fortune. This chapter explains the mechanisms that keep supply and demand in balance while coordinating production and consumption to produce the promised efficient outcome.

Chapter Summary 1-5: The plan of deregulation is to achieve efficiency through competition. Economics guarantees this result provided the market reaches a classic competitive equilibrium. This requires at least three conditions to be met: price taking suppliers, public knowledge of the market price, and well-behaved production costs. Although production costs seem problematic to many, they cause little trouble, and deregulation will probably succeed if markets are designed for maximum competition and transparent prices.

Section 1: Competition Means More than Struggle. The dictionary defines competition as "a struggle with others for victory or supremacy," but economics does not. Designing markets to be "competitive" in the dictionary's sports-oriented sense produces poor designs about which little can be predicted. Economic competition requires many competitors on each side of the market and results in a lack of market power and "price taking" behavior.

Section 2: Efficient Markets and the Invisible Hand. The central result of economics states that competition leads to efficiency. But to achieve short-run efficiency, competitive behavior must be supplemented with well-behaved costs and good information. Long-run efficiency requires free-entry of new competitors as well. Efficiency means that total surplus, the sum of profit and consumer surplus, is maximized.

Section 3: Short- and Long-Run Equilibrium Dynamics. Price and quantity adjustments, usually by suppliers, lead the market to equilibrium. In a competitive market, suppliers adjust output until marginal cost equals the market price and adjust price until the market clears (supply equals demand). They are price takers because when considering what quantity to produce they *take* the market price as given; that is, they assume it will remain unchanged if they change their output.

A long-run competitive equilibrium is brought about by investment in productive capacity. Profit (which means long-run economic profit) is revenue minus costs, and cost includes a normal return on capital (investment). Thus, zero economic profit provides a normal return on investment. If economic profit is positive and the market competitive, new suppliers will enter. In this way profit is brought down to zero under competition, but this is enough to cover all fixed costs and a normal risk premium.

Section 4: Why is Competition Good for Consumers? Competition minimizes long-run costs and pays suppliers only enough to cover these minimum costs. Although it is possible to depress price in the short run, it is not possible to pay less on average than minimum long-run average cost.

1-5.1 COMPETITION MEANS MORE THAN "STRUGGLE"

The dictionary defines **competition** as "a struggle with others for victory or supremacy." This definition is based on sports, not economics, but is quite influential with regulators and politicians. Consequently, when economics says "competition is desirable," this is often interpreted to mean that struggle among market players is desirable. There is a grain of truth to this interpretation, but it misses the main point.

The popular view judges competition mainly on fairness, so market power on the supply side is not a problem provided that demand is similarly endowed. Competition is now in vogue with many regulators, and many who have spent a lifetime passing judgment on the fairness of prices have taken up the call to "let the market do it." They see their new job as making sure the new markets are fair, that "the playing field is level." They believe it is only necessary to ensure the struggle between market players is fair. Because economics promises that competitive markets will be efficient, a good outcome is thought to be assured.

The economic promise of efficiency is not predicated on a fair struggle. Two fairly matched "competitors" do not approximate what economics means by competition. For example, economics makes no guarantee that pitting a monopoly transco against an equally powerful load aggregator will produce an even moderately acceptable outcome. Economics cannot predict the outcome of this kind of "competition" and would view this as a very poorly structured market.

The economist's notion of competition refers to competition among suppliers or among demanders but not between suppliers and demanders. Competition is not a struggle between those who want a higher price and those who want a lower price. The process of economic competition between many small suppliers works by suppliers undercutting each other's price in order to take away the others' customers. This drives the price down to the marginal cost of production but no lower because at lower prices suppliers would lose money. If supply-side competition is stiff enough, the market price will be pinned to the marginal-cost floor. This is the meaning of perfect competition.

When suppliers face such stiff competition that they cannot affect the market price and must simply accept it and sell all they can sell profitably at this price, they are said to be *price takers*. This is the principle requirement for a market to be perfectly competitive and is the primary assumption on which economic claims of market efficiency rest.

Generally it takes many competitors, none of which have a large market share, to produce perfect competition in the economic sense.[1] If there are any large suppliers they are likely to have the ability to profitably raise price. In this case they are not price takers and are said to have market power. They know they can affect the supply–demand balance by reducing their output and thereby drive up the price enough to increase their profit.

1. Under the special and uncommon conditions of Bertrand competition, two competitors are enough.

1-5.2 THE EFFICIENCY OF PERFECT COMPETITION

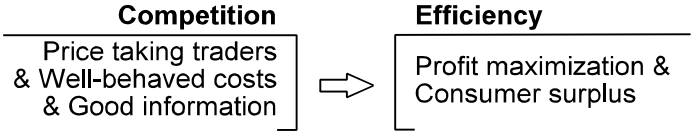
One economic result is, without doubt, the most prominent in all of economics. It is the point made by Adam Smith in the *Wealth of Nations:*

> *... he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention.*[2]

Vague as this may be, Adam Smith, and later Leon Walras, are correctly credited with developing the notion that competitive markets harness the profit motive to produce an efficient and socially useful outcome. This Efficient-Competition Result has been re-examined many times and modern proofs have resulted in "Nobel Prizes" for Kenneth Arrow (1972) and Gerard Debreu in (1983).[3]

Short-Run Competition

The Efficient-Competition Result has limitations. It does not mean that every free market is efficient, or even that every free market in which suppliers are price takers is efficient. Because of these limitations, economics has carefully defined both competition and efficiency and has added two more concepts: well-behaved cost functions and good information. The modern Efficient-Competition Result can be summarized as follows:



This result can also be summarized as "a competitive equilibrium is efficient."[4] The three conditions listed above under "competition," are necessary to guarantee that the market will reach a **competitive equilibrium**. If suppliers have small enough market shares, they will not have the power to change the market price and profit from doing so, and they will take price as given. This is called acting competitively, but it does not guarantee a competitive equilibrium. First, such an equilibrium must exist and second, traders must be able to find it. If costs are not well behaved—and startup and no-load costs are not—there will be no equilibrium. If traders lack adequate information, including publicly known prices, they may not

2. Smith is often quoted as saying that a market is "guided as if by an invisible hand." But a full text search of *The Wealth of Nations* reveals only this one use of "invisible hand." In his inaugural address of 1789, George Washington observed that the "Invisible Hand" (of God) had guided the United States to victory in 1776 (the same year Smith's book was published). In fact, both Smith and Washington viewed the invisible hand, God, and the forces of nature as being nearly synonymous.

3. Economists do not get authentic Nobel prizes. The prize in economics is given by the Bank of Sweden, not by the Nobel Foundation.

4. This discussion is necessarily far from rigorous and is meant only to convey a general understanding of the most important concepts. See Mas-Colell et al. (1995) starting on p. 308.

find the competitive equilibrium which consists of an optimal set of trades. (Problems with ill-behaved, **nonconvex**, costs and lack of information are discussed in Chapter 3-8.)

Definitions

Perfect Competition
Agents act competitively, have well-behaved costs and good information, and free entry is brings the economic profit level to zero.

Act Competitively
To take the market price as given (be a price taker).

Well-Behaved Costs
Short-run marginal cost increases with output and the average cost of production stops decreasing when a supplier's size reaches a moderate level.

Good Information
Market prices are publicly known.

Long-Run Competition

A short-run competitive equilibrium is (short-run) efficient; it makes the best use of presently available productive resources. A long-run competitive equilibrium guarantees that the right investments in productive capacity have been made but requires that the three short-run conditions be met and adds two new ones. Production costs must not possess the conditions for a natural monopoly (see Section 1-1.1), and competitors must be able to enter the market freely.[5] With **free entry**, if there are above-normal profits to be made, new suppliers will enter which will reduce the level of profits. In this way free entry ensures that profits will not be above normal. A normal profit level is the key characteristic of a long-run competitive equilibrium. **Barriers to entry** is the term used to describe market characteristics that prevent free entry.

Efficiency and Total Surplus

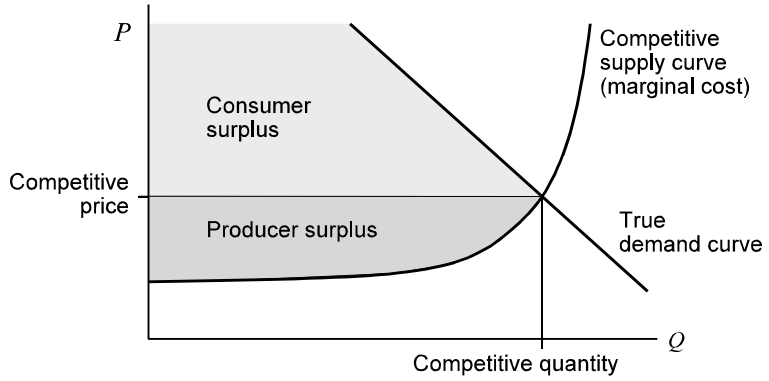
Almost every proposed market design is declared efficient, but in economics the term has a specific meaning. The simplest meaning applies to productive efficiency which means that what is being produced is being produced at the least possible cost. Minimizing cost is often the most difficult part of the market designer's problem, so this meaning is generally sufficient.

When not qualified as productive efficiency, efficiency includes both the supply and demand sides of the market.[6] **Efficiency** means (1) the output is produced by the cheapest suppliers, (2) it is consumed by those most willing to pay for it, and (3) the right amount is produced. These three can be combined into a single criterion by using the concept of **consumer surplus**.

5. See Mas-Colell et al. (1995, 334).

6. The term "allocative efficiency" is almost universally used to mean demand-side efficiency. But this is not the meaning found in economics dictionaries; which include both sides of the market and do not distinguish it from "efficiency."

Figure 1-5.2
Total surplus equals the area between the demand curve and the marginal cost curve.



Definitions

Productive Efficiency
Production costs have been minimized given total production.

Efficiency
Total surplus has been maximized. This automatically includes minimizing the cost of what is produced and maximizing the value of what is consumed, as well as producing and consuming the right amount.

A consumer's demand curve measures how much the consumer would pay for the first kilowatt-hour consumed, and the second, and so on. Generally the more consumed, the less would be paid for the next kilowatt-hour. Because the initial kilowatt-hours are so valuable, the total value of consumption is generally much greater than the amount paid. The difference between the maximum a consumer would pay as revealed by the consumer's demand curve and what the consumer actually does pay is the consumer's surplus.

Profit is analogous to consumer surplus and is often called **producer surplus**. It is total revenue minus total cost, while consumer surplus is total value to consumers, V, minus total consumer cost, CC. (V is sometimes called gross consumer surplus, and is the area under the demand curve for all consumption.) Both V and CC are measured in dollars. If the sum of profit and consumer surplus is maximized, the market is **efficient**, and all three of the above criteria follow.

When profit ($R - C$) and consumer surplus ($V - CC$) are added, the consumer costs (CC) and producer revenue (R) cancel because they are the same. The result is **total surplus** ($V - C$), consumer value minus producer cost. Consequently efficiency is the same as maximizing total surplus.

The Efficient-Competition Result

Result 1-5.1 **Competitive Prices Are Short- and Long-Run Efficient**
If productions costs are well behaved so competitive prices exist, these prices will induce short-run (dispatch) efficiency and long-run (investment) efficiency.

Note that the Efficient-Competition Result does not say that a market's prices will be competitive if costs are well behaved. That conclusion requires the lack of market power and good information. The Efficient-Competition Result says that if market prices are competitive then supply and demand will be efficient. Assuring competitive prices is the main problem addressed by this book.

Problems Caused by Production Costs

The Efficient-Competition Result depends on well-behaved production costs and these cannot be designed. Either efficient generators have well-behaved costs or they do not. If they do not, and costs are sufficiently problematic, then a standard competitive market design cannot be depended on to provide an efficient outcome.

Cost problems present the most fundamental challenge. Three different problems receive attention: (1) nonconvex operating costs, (2) the fixed costs of investment, and (3) total production costs that decrease up to very large scales of production. The third problem is the problem of natural monopoly and was discussed in Chapter 1-1.

The second problem is the subject of the most frequent misconception, an extremely pessimistic one. It holds that ordinary fixed costs are sufficient to disrupt a competitive market. This pessimism about competition is often accompanied by optimism about less-competitive free markets. The belief is that if the market can avoid the problems of a competitive equilibrium free-market forces will produce a good outcome. Perhaps, it is argued, if the market is not monitored too closely, generators will exercise market power and thereby earn enough to cover fixed costs and keep themselves in business. Chapter 2-1 shows that fixed costs are not a problem for competition and the proposed noncompetitive remedy is unnecessary and detrimental.

The problem of **nonconvex** operating costs is the most difficult. The cost of starting a generator makes generation costs nonconvex because it makes it cheaper per kWh to produce 2 kWh than to produce 1kWh. This causes the market to lack a competitive equilibrium and could easily cause inefficiency in the dispatch of an otherwise competitive market. To circumvent this problem, some markets use a unit-commitment auction that attempts to replace a standard "classic" competitive equilibrium with a different equilibrium which is still efficient. Chapter 3-9 discusses this problem in detail and suggests it may be of minor importance and that a standard competitive market might still provide a very high level of efficiency.

These conclusions are part of a larger pattern. Competitive markets are difficult to design, and none of the three basic requirements of efficiency can be achieved to perfection. But if the two controllable ones, price taking behavior and good information, are well approximated, a very efficient market will result. The quirks of a competitive equilibrium are not much of a problem, but designing a competitive market requires a great deal of care.

1-5.3 SHORT- AND LONG-RUN EQUILIBRIUM DYNAMICS

Markets are never in equilibrium, but economics focuses primarily on their equilibrium behavior. The ocean is never in equilibrium, yet it is always found at the lowest elevations where physics predicts its equilibrium to be. In equilibrium the ocean would have no waves. Although markets, like oceans, have "waves," they too usually stay near their equilibrium. An equilibrium may change over time as the globe warms and the ice caps melt, but this does not prove the equilibrium uninteresting. A market's equilibrium is a useful guide to its behavior, even though the market is never exactly in equilibrium.

The Meaning of Short Run and Long Run

These concepts do not, as is often supposed, refer to specific periods of time but instead refer to the completion of particular market adjustment processes. "In the short run" indicates that adjustments in the capital stock (the collection of power plants) are being ignored, but adjustments in the output of existing plants are being considered.

The phrase "in the long run" indicates adjustments in the capital stock are not only being considered but are assumed to have come to completion. This is a useful abstraction. If the market has not recently suffered an unexpected shock, it should be near a state of long-run equilibrium because business spends a great deal of effort attempting to discern future conditions, and for the last five years, today was the future.

Of course mistakes are made and markets are never in exact long-run equilibrium. But mistakes are as often optimistic as pessimistic, and consequently a long-run analysis is about right on average. However, a newly-created market is more likely than most to be far from its equilibrium.

Both the supply and the demand side of the market adjust their behavior in order to produce a market equilibrium, but competitive economics is primarily concerned with the supply side.[7] This section explores the forces that push the supply side of a market toward a competitive equilibrium.[8]

The Short-Run Equilibrium

Marginal cost is the cost of producing one more unit of output, one more kilowatt-hour. It is also approximately the savings from producing one less kWh. In this section and the next, these are assumed to be so close together that no distinction is necessary, which is typically the case. Chapter 1-6 pays a great deal of attention to the special case where these are different.

In a competitive market suppliers are **price takers**. They cannot change the market price profitably, so they consider it fixed. Price taking also means they can sell all they want at the market price, but they cannot sell anything at a higher price. Most markets are not perfectly competitive, and suppliers find that at a higher price they sell less but more than nothing. This will be ignored as the present purpose is to analyze how a market would work if it *were* perfectly competitive.

A short-run competitive equilibrium determines a market price and a market quantity traded. To bring the market into equilibrium, two dynamic adjustment mechanisms are needed: (1) a price adjustment and (2) a quantity adjustment. In most markets suppliers adjust both, although in some, buyers set the price.

7. Demand-side management concerns itself primarily with information problems on the demand side of the market. These problems also deserve attention.

8. For a more complete treatment of the microeconomics of competition presented with power markets in mind, see Rothwell and Gomez (2002).

Quantity Adjustment. A price taking supplier will *increase* output if its marginal cost, *MC*, is less than the market price, *P*, and will *decrease* its output if $MC > P$. Its profit increases by $(P - MC)$ for every unit produced when *P* is higher than *MC* and decreases by $(MC - P)$ when *P* is lower.

Price Adjustment. Whenever demand exceeds supply, suppliers raise their prices, and whenever supply exceeds demand, they lower prices.

Equilibrium. The quantity adjustment dynamic causes the marginal cost to equal the market price in a competitive market. The price adjustment dynamic causes the quantity supplied to equal the quantity demanded. When supply equals demand, the market is said to have *cleared*, and the price that accomplishes this is called the **market-clearing price**, or the equilibrium price, or, for a **competitive market**, the **competitive price**. Together the two adjustment mechanisms bring a competitive market to a **competitive equilibrium**.

Price Taking vs. Price Adjustment. Notice that "price taking" suppliers adjust their prices in order to clear the market. This is not a contradiction. "Price taking" is something that happens in the quantity-adjustment dynamic but not in the price-adjustment dynamic. Price takers "take the price as given when computing their profit-maximizing output quantity." This means they assume that their choice of output will not affect the price they receive for it.

The quantity dynamic, which causes *MC* to equal the market price, acts as a coordinating mechanism among suppliers because there is only one market price. This is why public knowledge of the market price is a key assumption of the Efficient-Competition Result. Because all suppliers have the same marginal cost in the competitive equilibrium, no money can be saved by having one produce more and another less. This is what makes production efficient.

Some will object to this result on the grounds that coal plants have lower marginal costs than gas turbines even in a competitive equilibrium. This objection is based on a misunderstanding of the definition of "marginal cost," which will be explained in the following chapter.

Price Taking and Price Adjusting

Suppliers typically name their price. For example, most retailers put price tags on their wares, and customers pay those prices. So how can suppliers be "price takers?"

Because "price-taking" has a specialized meaning, suppliers can be both price takers and price adjusters at the same time. Suppliers take price as given when deciding how much to produce and adjust their price if they notice excess supply or demand in the market.

The Marginal-Cost Pricing Result

Result 1-5.2 **Competitive Suppliers Set Output So That $MC = P$**

A competitive producer sets output to the level at which marginal cost equals the market price, whether or not that is the competitive price. This maximizes profit. ($MC = P$ for all suppliers.)

The Long-Run Equilibrium

The process of long-run competition involves investing in plant and equipment, not simply changing the output of existing plants. This dynamic requires a definition of profit. Profit is of course revenue minus cost, but economics defines costs more broadly than does business. Economics, and this book, define cost to include a normal rate of return on all investment. This rate of return is defined to include a **risk premium**. If a supplier covers its costs, it automatically earns a normal rate of return, including an appropriate risk premium, on its entire investment. Under this definition of "normal," a business that earns more is considered to be worth investing in, and a business that earns less is not. A normal business investment, therefore, has revenues that exactly cover all its costs in the economist's sense. Because **profit** equals revenue minus cost, a normally profitable supplier earns zero profit.

Definition

(Economic) Profit
Revenue minus total cost, where total cost includes a normal, risk-adjusted, return on investment. The normal (economic) profit level is zero. (Business defines a normal return on equity to be profit, while economics defines it as covering the cost of equity.)

Short-Run Profit
Revenues minus short-run costs which include variable, startup and no-load costs. The "profit function," defined in Chapter 2-7, computes short-run profits.

As defined, profit is synonymous with long-run profit which is different from short-run profit which does not include the cost of capital; that is, it does not include any return on investment. Consequently, short-run profit is expected to be positive on average so these profits can cover the fixed cost of capital.

Result 1-5.3a **Under Competition, Average Economic Profit Is Zero**
In a long-run competitive equilibrium, the possibility of entry and exit guarantees that profits will be normal, which is to say zero.

Result 1-5.3b **Under Competition, Fixed Costs Are Covered**
When profit is zero, all costs are covered including fixed costs, so in the long run, competition guarantees that fixed costs will be covered.

Result 1-5.3c **A Supplier with a Unique Advantage Can Do Better**
If a supplier has access to limited cheaper inputs (hydro-power or geothermal energy), it will have greater profits. If the advantage is unlimited, it has a natural monopoly.

If the expected market price is so low that a supplier cannot enter the market and cover all costs, no supplier will enter. More specifically, if a new generation unit cannot cover all costs, no new units will be built. The result will be a gradually diminishing supply of generation (due to retirements of old plants) in the face of

gradually increasing demand. This tightening of the market will cause the price to rise, and eventually price will be high enough to cover all costs.

Similarly, if price is so high that costs are more than covered, suppliers will build new generating units. This will increase supply and cause the price to fall. The result of this long-run dynamic is that the profit in any competitive market returns to the normal level of profit (zero) in the long-run competitive equilibrium.

1-5.4 WHY IS COMPETITION GOOD FOR CONSUMERS?

In the long-run producers cover their fixed costs, and in the short run total surplus is maximized, but what consumers want is a low price. Does competition provide the lowest possible price?

Not in the short run. In the short run, it is possible to design market rules which lower the market price without reducing supply. This is difficult but possible. But at a lower price producers will not cover their fixed costs. This will make future investors think twice. The result will be a risk-premium added to the cost of capital and future production will be more costly than it would have been had cost been left at the competitive level.

Competition does not guarantee the lowest possible price at any point in time. Instead it guarantees that suppliers will just cover the long-run total costs and no more. It also guarantees that the cheapest suppliers will be the ones producing. Together these mean production costs (including the long-run cost of invested capital) are minimized and producers are paid only enough to cover their cost. This implies that the long-run average cost to consumers is also minimized. No market design regulated or unregulated can induce suppliers to sell below cost on average. Competition minimizes long-run average costs of production and long-run average costs to consumers.

Chapter 1-6

Marginal Cost in a Power Market

The trouble with the world is not that people know too little,
but that they know so many things that ain't so.

Mark Twain
(1835–1910)

SIMPLIFIED DIAGRAMS OF GENERATION SUPPLY CURVES HAVE CONFUSED THE DISCUSSION OF MARGINAL COST. Typically, these supply curves are diagrammed to show a constant marginal cost up to the point of maximum generation. Then marginal cost becomes infinite without taking on intermediate values. Typically it jumps from about $30 to infinity with only an infinitesimal increase in output. Mathematics calls such a jump a discontinuity. In fact, the curve would be discontinuous if it jumped only from $30 to $40.

The standard definition of marginal cost does not apply to the points of discontinuity. Hence it does not apply to a right-angle supply curve at the point of full output, neither does it apply to the points of a market supply curve at which it jumps from one generator's marginal cost to the next. Unfortunately market equilibria sometimes occur at such points, and concerns over market power often focus on them. Attempts to apply the standard definition at these points can produce confusing and erroneous results.

Fortunately, the definition is based on mathematics that generalizes naturally to discontinuous curves. Applying this generalization to the textbook definition clears up the confusion and restores the economic results that otherwise appear to fail in power markets. For example, in power markets, as in all other markets, the competitive price is never greater than the marginal cost of production.

Chapter Summary 1-6: Individual supply curves are often constructed with an abrupt end that causes the market supply curve to have abrupt steps. The standard marginal-cost definition does not apply at such points. Instead, left- and right-hand marginal costs should be used to define the marginal-cost range. Then the competitive price, which remains well defined, will always lie within that range. A market price exceeding the marginal-cost range indicates market power.

Section 1: The Role of Marginal Cost. Marginal costs play a key role in cost-based power auctions because they help determine the competitive price. They also play a key role in analyzing market power and gain their importance by defining the competitive supply curve for individual generators. To find the market (aggregate) supply curve, individual supply curves are summed horizontally.

Section 2: Marginal-Cost Fallacies. In power-market analysis, marginal cost is often defined as the cost of the last unit produced, but this definition is found in no economics text. A second fallacy asserts that when marginal cost is ambiguous, the competitive price is ambiguous. Together these lead to a variety of erroneous conclusions, such as "the competitive price is above marginal cost," and "the competitive price is ambiguous."

Section 3: The Definition of Marginal Cost. When a marginal-cost curve is discontinuous (has a sudden jump), marginal cost can be specified only within a range at the points of discontinuity. This range extends from the left-hand to the right-hand marginal cost at the point under consideration. For all points where the curve is continuous, the range is a single point equal to the standard marginal cost.

Section 4: Marginal Cost Results. The competitive price is within the marginal-cost range of every competitive generator and within the marginal-cost range of the market. If even one supplier has a supply curve that is continuous at the market price, the market supply curve is continuous at that price and the competitive price is equal to the standard marginal cost which is well defined. In any case, the competitive price is the price at which the supply and demand curves intersect.

Section 5: Working with Marginal Costs. This book assumes that supply curves have extremely large but finite slopes rather than the infinite slopes frequently assumed. This is a more realistic assumption and has no practical consequences, but it has the simplifying property of making marginal cost well defined and the marginal cost of all operating competitive generators equal to the market price.

Section 6: Scarcity Rent. Scarcity rent is revenue less variable cost and is needed to cover startup and fixed costs. Economics refers to this as "inframarginal rent," and has no separate definition of a scarcity rent. A folk-definition defines scarcity rent as actual revenue minus the maximum revenue that is collected just before the system runs completely out of capacity. Used with a stylized model, this definition has some appeal, but when applied to real systems it is highly ambiguous and misleading.

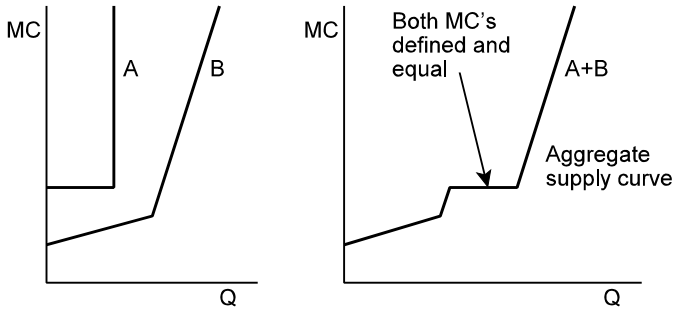


Figure 1-6.1
Adding individual supply curves horizontally to find the market supply curve. If B is continuous, A + B is also.

1-6.1 THE ROLE OF MARGINAL COST

Marginal cost plays a key role in the economic theory that proves a competitive market is efficient, but there are also two practical uses of marginal cost that increase its importance in a power market. First, many power markets rely on a central day-ahead auction in which generators submit individual supply curves and the system operator uses these to determine the market price. Because price should equal marginal cost in an efficient market, the auction rules should be informed by a coherent theory of marginal cost. Second, many power markets suffer from potential market-power problems which cause the market price to diverge from marginal cost. Market monitors need to understand this divergence.

Although the competitive market price usually equals the marginal cost of production, it is not determined by that alone. At times marginal cost is ambiguous, yet the competitive price is not. Then, marginal value (to customers) plays the decisive role. The competitive price is determined by the intersection of the market's supply and demand curves. Marginal cost determines only the supply curve.

A supply curve can be thought of as answering the question, How much would a generator produce if the market price were $\$P$/MWh? As explained in Section 1-5.3, price-taking suppliers adjust output until marginal cost equals the market price. As a consequence, if Q is the quantity supplied at a given price P, then P must equal the marginal cost. Thus a price-taker's supply curve and marginal cost curve are the same.

The market's supply curve, also called the **aggregate supply curve**, is found by summing horizontally all of the individual generators' supply curves. For a given price, the quantity supplied by each generator is read horizontally from each individual supply curve and these quantities are summed to find the market supply. This quantity is plotted at the given price, as shown in Figure 1-6.1.

Notice that because one generator has a continuous supply curve (no vertical section) the market has a continuous supply curve. Notice also that when both generators are operating and have defined marginal costs, they have the same marginal cost. Section 1-6.3 generalizes this by showing that every operating generator either has a marginal cost equal to the market price or has a marginal-cost range that includes the market price.

1-6.2 MARGINAL-COST FALLACIES

Discontinuous Supply Curves

Individual supply curves are almost always drawn as "hockey sticks." That is, they are drawn with a slight upward slope (or as flat) until they reach the capacity limit of the generator and then they are drawn as perfectly vertical (see curve A, Figure 1-6.1). Textbook supply curves usually have a slope that increases gradually (See curve B, Figure 1-6.1). Curves without a vertical segment are called continuous. Unfortunately, a generator's supply curve, as typically drawn, takes an infinite upward leap when it reaches full output (which is the most common output level for an operating generator). At this point, marginal cost is not smooth but jumps from say \$30/MWh to infinity with only an infinitesimal change in output.

The smoothness of textbook supply curves plays a crucial role in keeping the textbook definition of marginal cost simple, and this has led to mistakes and confusion. Eliminating the confusion requires the introduction of a carefully constructed definition which applies to the discontinuous supply curves used in power-market analysis. With this definition of marginal cost, all standard economic results are found to apply to power markets. Once this is understood, the problematic supply curves can be analyzed correctly with a simple rule of thumb. This provides guidance when setting the market price in a cost-base auction and when determining whether market power has been exercised.

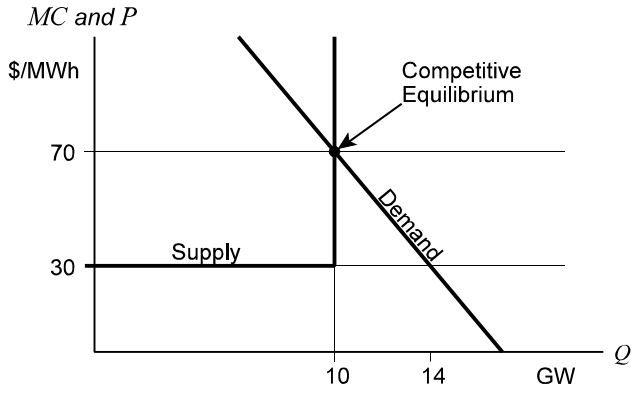
Fallacies

Two basic fallacies underlie a series of misconceptions surrounding competitive pricing and market power. These are (1) the Marginal-Cost Fallacy and (2) the Ambiguous-Price Fallacy. Both of these will be illustrated using Figure 1-6.2, which shows a normal demand curve and a supply curve that is constant at \$30/MWh up to an output of 10 GW, the capacity limit of all available generation.

The Marginal-Cost Fallacy takes two forms. The simple form asserts that marginal cost at Q = 10,000 MW is \$30/MWh in Figure 1-6.2. The subtle form asserts that nothing can be said about the marginal cost at this output level. Some of the conclusions drawn from these assertions are as follows:

1. The competitive price is \$30/MWh, and the market should be designed to hold prices down to this level.
2. The competitive price is \$30/MWh, and this is too low to cover fixed costs, so marginal-cost prices are inappropriate for power markets.
3. Scarcity rents are needed to raise prices above marginal-cost-based prices.
4. Market power is necessary to raise prices to an appropriate level.
5. The competitive price cannot be determined.

Figure 1-6.2
A normal market equilibrium for an abnormal supply curve.



All of these conclusions assume that there is some problem with standard economics caused by the supply curve coming to an abrupt end instead of turning up smoothly as it does in undergraduate texts. In fact, economic theory has no difficulty with this example, and all of the above conclusions are false. Consider a competitive market, with many suppliers and many customers, described by the curves in Figure 1-6.2. What if the price in this market were \$30/MWh? At this price, the demand curve shows an excess demand of about 4 GW. Some customers trying to buy more power are willing to pay up to \$70/MWh for another MW of supply. They will find a supplier and offer to pay considerably more than \$30, and the supplier will accept. This shows that the competitive price is above \$30/MWh. The story will be repeated many times, with different values, until the market price reaches \$70/MWh. At that price every supplier will produce at full output, so the supply will be 10 GW, and demand will be 10 GW. At any higher price demand would fall short of supply, so the price would fall, and at any lower price, demand would exceed supply, so the price would rise. There is nothing unusual about this equilibrium; it is the classic story of how price clears a market by equating supply and demand.

The Marginal-Cost Fallacy

Fallacy **1-6.1** **Marginal Cost Equals the Cost of the Last Unit Produced**

Marginal cost equals the savings from producing less even when this is different from the cost of producing more.

(Subtle Version)

Nothing can be said about marginal cost at the point where a supply curve ends or jumps from one level to another.

But shouldn't price equal marginal cost? In this example, all that can be said is that marginal cost is greater than \$30/MWh. So there is no contradiction between price and marginal cost, but they cannot be proven to be equal. The desire to pin down marginal cost precisely seems to arise from a belief that competitive suppliers should set price equal to marginal cost and thereby determine the market price. But this logic is backwards. As explained in Section 1-5.3, suppliers set price to

clear the market and set quantity to bring marginal cost in line with price. In this example, the market-clearing forces of supply and demand determine price unambiguously, and although marginal cost is ambiguous, it is greater than $30/MWh which is enough to determine supply unambiguously. Everything of practical importance is precisely determined.

The Ambiguous-Price Fallacy

Fallacy **1-6.2** **When Marginal Cost Is Ambiguous, so Is the Competitive Price**
Competitive suppliers set price equal to marginal cost; thus when marginal cost is hard to determine, the competitive price is hard to determine.

Having analyzed the example, the preceding list of incorrect conclusions can be restated correctly as follows:

1. The competitive price is not $30/MWh, and the market design should not hold price to this level.
2. The competitive price is high enough to contribute significantly to fixed cost recovery.
3. No mysterious "scarcity rent" need be added to the marginal cost of physical production.
4. Market power is not needed if the market is allowed to clear.
5. The competitive price is $70/MWh.

1-6.3 THE DEFINITION OF MARGINAL COST

The above discussion is accurate but informal. Because of the controversy in this area, it is helpful to formalize the concepts used in analyzing supply curves with discontinuities or abrupt terminations.

The *MIT Dictionary of Modern Economics* (1992) defines marginal cost as "the extra cost of producing an extra unit of output." Paul Samuelson (1973, 451) defines marginal cost more cautiously as the "cost of producing one extra unit more (or less)." The "or less" is important. The assumption behind this definition is that producing one more unit of output would cost exactly as much as producing one less unit would save. This is true for the continuous marginal-cost curves of textbook economics but not for the discontinuous curves used by power-market analysts. To discuss the marginal cost of a discontinuous supply curve, the definition must be extended to include the points of discontinuity where the cost to produce an extra unit is distinctly greater than the savings from producing one less.

Left- and Right-Hand Marginal Costs

In the example of Figure 1-6.2, the marginal cost of production goes from $30 on the left of 10 GW to infinity on the right of 10 GW. This is a double complication. Not only does marginal cost change abruptly, it becomes infinite. The present

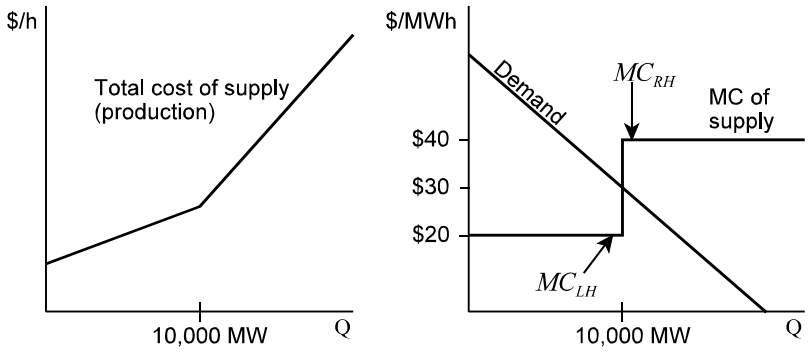


Figure 1-6.3
Right- and left-hand marginal costs.

definitions can be illustrated more clearly with a less pathological marginal-cost curve.

Figure 1-6.3 shows the total cost curve and the marginal cost curve of a simple market. The discontinuity is the jump in marginal cost at the 10 GW output level. To the left of 10 GW the marginal cost is \$20/MWh, while to the right it is \$40/MWh. But what is the marginal cost precisely at 10 GW? It is undefined, but, as every textbook would confirm, the answer is *not MC* = \$20/MWh.

To formalize this definition, it is useful to consider the mathematics of the total cost curve shown at the left of Figure 1-6.3. To the left of 10 GW, its **derivative** (slope) is \$20/MWh, while to the right its slope is \$40/MWh. But the mathematical definition of a derivative breaks down at 10 GW, and since marginal cost is just the derivative of total cost, the definition of marginal cost also breaks down at this point. Mathematics *does* define two very useful quantities at the 10-GW point, the left-hand derivative (slope) and the right-hand derivative (Courant 1937, 199–201). These are, of course, \$20 and \$40/MWh, respectively. Because marginal cost is just the derivative, it is natural to define **left-hand marginal cost** (MC_{LH}) as the left-hand derivative, and **right-hand marginal cost** (MC_{RH}) as the right-hand derivative. Other points along the total cost curve also have left and right-hand derivatives, and these are just equal to the normal derivative. Similarly, MC_{LH} and MC_{RH} are normally equal to each other and equal to standard marginal cost, *MC*.

The **marginal-cost range**, MC_R, is defined as the range of values between and including MC_{LH} and MC_{RH}. This definition is motivated by the idea that marginal cost cannot be pinned down at a point of discontinuity but can reasonably be said to lie somewhere between the savings from producing one less and the cost of producing one more unit of output.

Definitions

Left-hand marginal cost (MC_{LH})
The savings from producing one less unit of output.

Right-hand marginal cost (MC_{RH})
The cost of producing one more unit of output. When this is impossible, MC_{RH} equals infinity.

The marginal-cost range (MC_R)
The set of values between and including MC_{LH} and MC_{RH}.

1-6.4 MARGINAL COST RESULTS

Refining the Marginal-Cost Pricing Result

In Figure 1-6.2, the MC_{LH} at 10 GW is \$30/MWh, but what is the MC_{RH}? It is tempting to say it is undefined, but again mathematics provides a more useful answer. The MC_{RH} at 10 GW is infinite. This definition is both mathematically sound and useful because it allows a simple rewriting of the standard economic results concerning marginal costs.

Result 1-6.1 **Competitive Suppliers Set Output so $MC_{LH} \leq P \leq MC_{RH}$**
A competitive producer sets output to a level at which its marginal-cost range, MC_R, contains the market price, P, whether or not that is the competitive price.

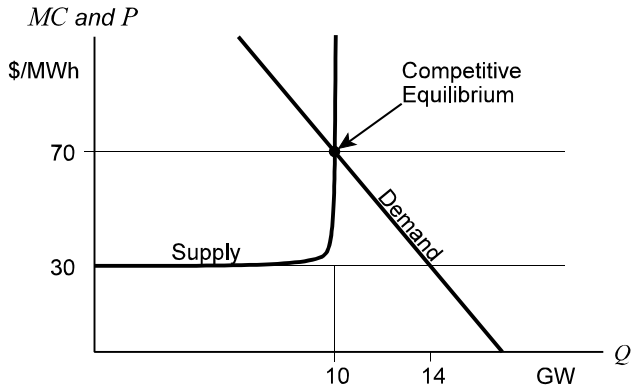
First, a price-taking supplier will decrease output as long as $P < MC_{LH}$ because producing one less unit will save MC_{LH} and cost only P in lost revenues. Thus, the savings is greater than the cost. Similarly, if $MC_{RH} < P$, the supplier will increase output. Thus whenever P lies outside the range between left- and right-hand marginal costs, the supplier will adjust output. When the range is below P, output is increased, which raises the range and vice versa when MC_R is above P. As a result, the marginal-cost range will end up encompassing P.

This means that in a competitive market, price will never exceed marginal cost; this would violate basic economics. Technically, $P > MC$ can never be proven true in a competitive market.[1] Competitive price will always be less than or equal to left-hand marginal cost, and there is no need for it exceed this value for fixed cost recovery.[2]

1. Even those who best understand these concepts sometimes add to the confusion. "*Thus in the absence of market power by any seller in the market, price may still exceed the marginal production costs of all facilities producing output in the market at that time.*" (Borenstein 1999, 3) ". . . the price of electricity has to rise above its short-run marginal cost from time to time, or peaking capacity would never cover its fixed costs." (Green 1998, 4).

2. Part 3 discusses "nonconvex costs," complexities of the production cost function that require deviations from marginal cost. Essentially this means that startup costs and other short-run, avoidable costs must be covered by price.

Figure 1-6.4
The smallest possible change in the supply curve of Figure 1-6.2 restores all normal economic properties.



This result can be extended from a single producer to the whole market. The MC_{RH} of the market is the least cost of producing one more unit, so it is the minimum of the individual marginal-cost ranges. Similarly, the market MC_{LH} is the maximum individual MC_{LH}. In a competitive market, every supplier is a price taker and adjusts its output until P is within its marginal-cost range. Thus P is less than or equal to every individual MC_{RH}, so it is less than or equal to the MC_{RH} of the supplier with the lowest MC_{RH}, which is the MC_{RH} of the market. Similarly, P is greater than or equal to the MC_{LH} of the market.

The range from the market MC_{LH} to the market MC_{RH} is contained within the marginal-cost ranges of each individual supplier. If even one supplier in the market has $MC_{LH} = MC_{RH}$, the market will also have this property. In other words, if even one supplier has a well defined-marginal cost at the market price, then the market itself has a well-defined marginal cost.

The System-Marginal-Cost Pricing Result

Result 1-6.2 **Competitive Price Equals System Marginal Cost**

In a competitive market, price is within the marginal-cost range of every generator supplying power. It is thus within the market's marginal-cost range. If even one operating supplier has a continuous marginal cost curve, the competitive price actually equals marginal cost as defined by the aggregate supply curve.

Finding the Competitive Price

Fortunately the above results are needed only for untangling the current confusions over marginal cost. They demonstrate, among other things, that price does not exceed marginal cost in a competitive power market.

Fortunately, these results are not needed to find the competitive equilibrium, which is determined, as in any other market, by the intersection of the supply and demand curves. This is most easily seen by smoothing out one of the problematic supply curves very slightly.

Standard economic theory applies once the vertical segments have been removed from the cost curves. This can be done with an arbitrarily small change in its shape.

As shown in Figure 1-6.4, giving the marginal cost curve a nearly, but not perfectly, vertical slope makes no noticeable difference to any economic result. And this is how it should be. Economics should not and does not depend on splitting hairs. Notice that in the finitely-sloped model, price really does equal marginal cost at the intersection of the two curves. The price and quantity dynamics of a market with the vertical supply curve will be essentially the same as those of the continuous market. In this example, at a price different from $70 and a quantity different from 10 GW, the markets have essentially identical gaps between supply and demand and between price and marginal cost. So they adjust price and quantity in the same way.

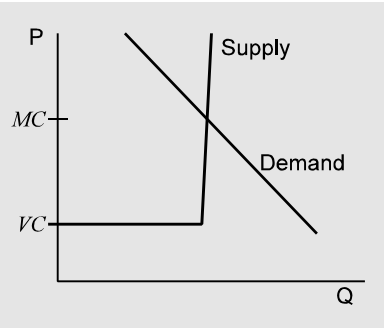
Result 1-6.3 **Supply Intersects Demand at the Competitive Price**
To find the competitive price and the marginal cost, draw the supply and demand curves, including the vertical parts of the supply, curve if any. The intersection of supply and demand determines "*MC*," *P*, and *Q*.

This demonstrates that the standard method of finding the competitive equilibrium works even when the marginal cost curves have infinite slopes. Of course if the slope is infinite at the intersection of supply and demand, marginal cost will be technically undefined. Yet pretending that the true marginal cost is determined by this simple short cut will never give the wrong answer to any real-world question.

1-6.5 WORKING WITH MARGINAL COSTS

Discussing left- and right-hand marginal costs and the marginal-cost range is cumbersome and unnecessary. If every vertical segment of a marginal cost curve is replaced with a nonvertical but extremely steep segment, the new curve will be continuous and will not jump from one value to another. Such a change may or may not improve its accuracy, but in either case it will make no detectable difference to any economic prediction of consequence.

This book will tacitly assume all supply curves and marginal cost curves that are depicted as having vertical segments actually have extremely steep but finite slopes. In other words, all marginal-cost curves are assumed to be continuous. Consequently, marginal cost is always a well-defined single value.



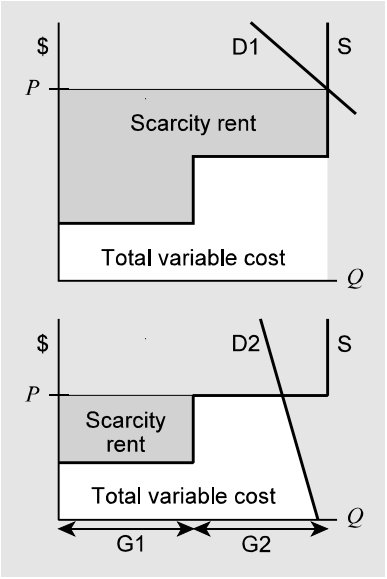
For example, a supply curve that is constant at $30/MWh up to a maximum output of 500 MW can be replaced with one that is identical up to 500 MW and then slopes upward linearly reaching a value of $30,000/MWh at an output of 500.001 MW. No measurement, however careful, could discern the difference. Yet this supply curve, being continuous, has a well-defined value (marginal cost) at every level of output.

In fact most, if not all generators, have continuous marginal cost curves. Typically, they have an **emergency operating range** above their nominal maximum output level and are willing to produce in this region if well paid or coerced. Most generators in PJM include such

an emergency operating range in their bids and the total capacity available in this range is 1,900 MW out of a total installed capacity of about 60,000 MW.[3] As long as there is one such generator in a market, the market's marginal cost curve is continuous. Real markets always have well-defined marginal costs and the competitive price equals that marginal cost. The difficulties resolved in this chapter only matter for the "simplified" diagrams used by power-market analysts.

This book also will use the same simplified diagrams but without taking the vertical segments literally. Such supply curves will have constant marginal costs up to the nominal "maximum" output level, but above that marginal costs will increase rapidly. If the supply curve is flat at $30 but the market price is $50, the generator's marginal cost will be $50 and it will produce on the steeply sloped segment. When referring to such a generator, it is both wrong and confusing to say its marginal cost is $30 as is the custom. To avoid this confusion, the marginal cost of a generator's supply curve to the left of the "maximum" output level will be termed its **variable cost**. This is not entirely standard, but it is in keeping with the term's normal usage which refers to all costs that vary with the output level.

1-6.6 SCARCITY RENT



"Scarcity rent" has no formal economic definition but many popular meanings.[4] Although several are useful, most do not lend themselves to careful analysis. However, one essential economic concept comes close to the popular meaning. **Scarcity rent** will be defined as revenue minus variable cost.[5] Economics refers to scarcity rent as **inframarginal rent**.

In the figure at the left, when demand is described by D1, both generators are producing at full output, and load would be willing to pay either generator more than its variable cost of production if it would produce more. In this sense they are both scarce and both earn scarcity rents.

With demand reduced to D2, as shown in the lower half of the figure, generators of type G2 have excess capacity and are no longer scarce and earn no scarcity rent; their variable costs equal the market price. Generators of type G1 are still scarce because load would be more than willing to pay their variable cost if they would produce more. If G2 had a variable cost of $1,000/MWh so that G1 were earning a rent of, say, $950/MWh because G1 could not satisfy the entire load, G1 would commonly be seen as in scarce supply. The above definition coincides with an important concept of economics and with the common meaning of scarcity.

3. Personal communication from Joe Bowring, head of PJM's Market Monitoring Unit, January 7, 2002.

4. Samuelson (1973, 623) comes close to using the term when he says "Competitively determined rents are the results of a natural scarcity." His definition of such rents is the long-run analog of the short-run definition of scarcity rent given here.

5. This is greater than short-run profit by the amount of startup costs and no-load costs which will be ignored until Part 3.

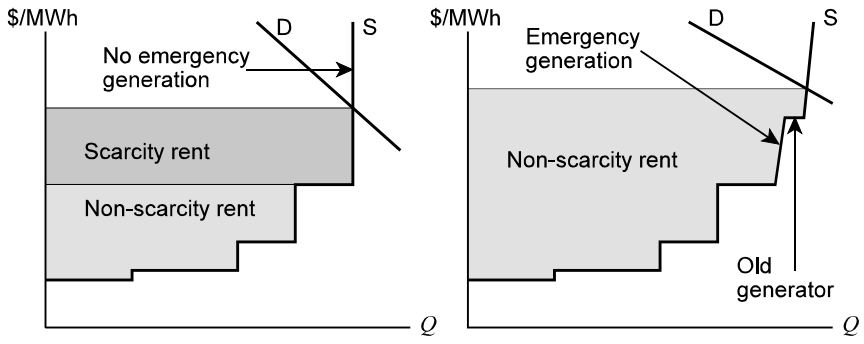


Figure 1-6.5
Folk-definition of scarcity rent.

A Folk Definition of Scarcity Rent

Sometimes, in the field of power system economics, scarcity rent is defined as actual revenue less the highest revenue earned before total generation becomes scarce. This might be called a "folk definition." The notion is that until the system runs out of capacity, price increases are due to increases in marginal cost, but after that point they are driven up by ever increasing scarcity. In an *idealized model*, this definition has some appeal.

Definition	**Scarcity Rent** Revenue minus variable operating cost (which do not include startup costs and no-load costs).

Say there are only ten types of generators on the market, and call the one with the highest variable cost the **peaker**. Next assume that there are no out-of-date generators with higher variable costs installed in the system. Finally assume that no installed generator has an **emergency operating range** in which its marginal costs increase dramatically as it increases its output beyond its normal rating. With these assumptions, peakers will earn enough to cover more than variable cost only when the system runs out of capacity. In other words, peakers can cover their fixed costs only from scarcity rents but not from any nonscarcity inframarginal rents. All other generators cover their fixed costs from a combination of scarcity and nonscarcity rents. The left half of Figure 1-6.5 illustrates this property of an idealized supply curve.

The folk definition has the advantage of allowing the following types of statements which seem designed to segregate scarcity conditions from the normal operating conditions of the market.

1. Scarcity rents pay capital costs of units that run infrequently.
2. In the long-run competitive equilibrium, scarcity rents are just high enough to cover the fixed costs of peakers.
3. Scarcity rents are paid only infrequently.

This appears to ratify the view that power markets are qualitatively different in their cost structure and consequently cannot be analyzed with the standard marginal-cost apparatus.

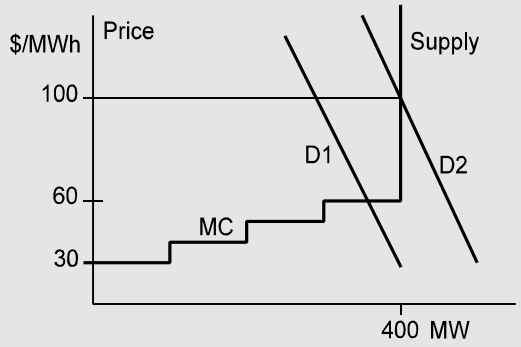
In the idealized model, these statements are true, although they give the impression that scarcity rents are mainly or wholly associated with peakers. In fact, under the folk definition, every type of generator receives the same amount of scarcity rent per MWh. In addition, the average scarcity rent in $/MWh does not equal the fixed cost of peakers but is greater by a factor of one over the duration of the peaker's use, something that is not easily determined.

Two problems with this definition make it unworkable in a real market. First, there are likely to be old generators on the system with variable costs greater than the most expensive new generator that would be built (the peaker). In this case scarcity will not set in until the old generator is at full output. This will expand the nonscarcity rents and shrink the scarcity rents to the point where they no longer cover the fixed costs of a peaker. Second, there will be some (probably many) generators with marginal cost curves that continue on up to some very high but ill-defined value. This will reduce scarcity rents to some negligible and indeterminable value. Proving scarcity rents exist requires proving price is above the point where the supply curve becomes absolutely vertical; absolutes are notoriously hard to prove.

Because of these shortcomings and the limited usefulness of the folk definition, this book will use only the definition given above that coincides with "inframarginal rents," a term that has proven itself useful in economics. This is in keeping with the chapter's general view that generation cost functions present no new problems of consequence and require only a minimal expansion of the definition of marginal cost and then only to deal with the stylized mathematics of discontinuous cost functions.

A Marginal-Cost Example

- Four suppliers can each produce 100 MW but no more.
- Each supplier has constant marginal cost (*MC*) up to this limit.
- Marginal costs and demand are as shown in the figure.



If demand is given by D1,

1. The competitive price is $60/MWh.
2. Any higher price indicates market power.
3. If the market is competitive, no supplier has $MC < \$60/MWh$.

If demand is given by D2 and the suppliers are price takers,

1. The market price (P) will be $100/MWh.
2. No generator will have a marginal cost of less than $100/MWh.
3. No market power is exercised at this price.
4. P is greater than the cost of the last unit produced ($60/MWh).

In both cases the marginal-cost rule for competition is

$$MC_{LH} \le P \le MC_{RH} \text{ *}$$

This is sufficient to determine the competitive market price and output.

* MC_{LH} is the savings from producing one unit less. MC_{RH} is the cost of producing one unit more and is considered arbitrarily high, or infinite, if another unit cannot be produced.