Part Two

BENEFITS OF DISTRIBUTED RESOURCES

2.1 Introduction

For the reasons and as part of the historical processes described in Part One, market actors choosing from the electrical resource menu summarized in Section 1.2.2 are undergoing a radical shift from a short menu of the most centralized resources toward a large and diverse menu favoring more appropriate scale. A simple, though partial, explanation of this shift is the desire to minimize regret[1]—either at what one did that one wishes one hadn't done, or at what one didn't do that one wishes one had done.

In a world of increasingly rapid technological and social change, minimizing regret is greatly aided by picking options that are relatively small, fast, modular, and cheap. Sections 2.2, "System Planning," and 2.3, "Construction and Operation," describe how this way of managing risk so as to minimize regret can yield important and measurable economic benefits. Subsequent sections describe distributed benefits related to T&D (the grid); to system operation; to the quality of electrical services provided; and to social and environmental factors. Implications of these principles, barriers to their adoption, and recommendations for further action are then surveyed in Part Three.

We are now ready to explore these approximately 207 kinds of distributed benefits as systematically as current understanding and published results allow. However, three general caveats are important first:

1. *The total value of distributed benefits depends strongly on technology- and site-specific details.*
2. The total value also depends on *which benefits are counted.* In general, assessments that find relatively modest gains from counting distributed benefits, such as one 1994 survey's 4–46% gain (over central-station generation) for photovoltaics or 2–78% for wind (54), omit many significant classes of benefits. A basic lesson of Part Two will be that the harder you look, the more distributed benefits you are likely to find, and that though many of those benefits are individually small, they are so numerous that they can still be collectively large.
3. Because such limited resources have been applied to codifying and quantifying distributed benefits, the explanations and evidence we can present, especially on how much each benefit is worth, *vary widely in type* (estimates, formal calculations, field examples, etc.); in *application to particular places, systems, and times*; and in their *accuracy and rigor*.

It is not yet possible to present a neat package of analytic solutions, practical examples, lookup tables, and the rest of the toolkit that a planner would like to take off the shelf and apply. The art and science of understanding distributed benefits are far too immature for that—certainly in the open literature, and probably also even if all the proprietary literature were available. However, we have presented summary boxes and other guideposts to help clarify the relationship of the different benefits; and to avoid cluttering the narrative flow with tutorials, definitions, examples, and technical notes, we have boxed these separately as labeled sidebars.

We can only hope that this assemblage of descriptions and examples, from many disparate places and with often wildly differing

[1] This valuable phrase was coined by Group Planning at Royal Dutch/Shell in London.

levels of detail and precision, will stimulate others with greater skills and resources to expand and refine this exploration with the level of effort it merits. We trust that such improvements will focus on putting the greatest care into refining the precision of the most valuable terms, rather than seeking spurious or needless precision in unimportant terms—mindful of Aristotle's terse admonition that in addressing any problem, educated people "seek only so much precision as its nature permits or its solution requires." (13)

2.2 System Planning

A noted text on corporate decision-making, *The Management of Scale: Big Organizations, Big Technologies, Big Mistakes* (138), examines case studies of disastrous large-scale blunders. Among their central causes, it identifies the adoption of inflexible technologies—those with "long lead time, large unit size, dependence upon infrastructure[,] and capital intensity." (139) Such a technology has the further attributes that:

> (a) Its development is to the direct benefit of large business organizations, able to spread some of the risk into public pockets.
>
> (b) It is likely to be an expensive failure.
>
> (c) Decision-making is highly centralized, with little debate, excluding some groups that are deeply affected by the technology.
>
> (d) The technology could have been identified as inflexible very early in its life.
>
> (e) More flexible technical alternatives exist.
>
> (f) These alternatives could be developed by organizations that are less centralized.

Many electric utilities bear extensive financial and psychological scar-tissue from their encounters with such technologies, particularly nuclear power. But as Part One described, among the key drivers of those multi-hundred-billion-dollar commitments were the perceptions that the giant plants would be necessary to keep the lights on and that they would decrease $/kW capital cost, presumed to be a surrogate for the cost of electric services. A critical part of the unraveling of this dogma was the realization that the hoped-for economies of scale were illusory and that a more sophisticated view of total cost and risk could even favor smaller units.

Rare wisps of internal criticism emanated from within the utility industry starting around 1970, but few if any squarely addressed the risks of gigantism; most, like those of Philip Sporn, dealt instead with demand forecasts and the balance between nuclear and fossil-fueled technologies (78, 297). Among the first wide cracks in the façade to be supported by rigorous analysis came in 1978, when John C. Fisher of the General Electric Company published a toned-down analysis through EPRI, and a more outspoken version in an international symposium, that was among the industry's first expert and explicit acknowledgements of diseconomies of unit scale.

Fisher presented a multiple-regression analysis of about 750 fossil-fueled steam power stations entering U.S. service during 1958–77 (238). He concluded, as he summarized in a letter (239), that

> Units with larger ratings take longer to build[,] and cost more on that account; units with larger ratings break down more often and take longer to repair and hence are out of service a larger fraction of the time. Because construction is slowed [*sic*] for larger units, the anticipated construction scale economy is diminished. Because

> reliability falls off for larger units[,] the anticipated operational scale economy is reversed for units larger than an optimum size. When the cost [reductions]...associated with replication of standardized units are recognized, the optimum size shrinks to the smallest possible size consistent with maintaining full performance quality for whatever technology is being employed. For subcritical fossil steam units (the most common utility central station steam unit)[,] this size is in the neighborhood of 125 MW....

That size was only *one-tenth* the maximum then being ordered, but was consistent with British findings that estimated a 200–300-MW optimum taking fewer factors into account (1). Taking qualitative account of flexible siting, reduced reserve margin, and perhaps smaller maintenance staffs because of higher unit reliability, the conclusion drawn—heretical then, but prescient in light of GE's and other firms' later success with combined-cycle gas turbines—was:

> The replication of a series of identical generating units opens up an entirely new and profoundly different avenue for reducing the capital cost of generating capacity. The economy of scale assumes a new form, and manifests itself as the reduction of cost that can be achieved through the scale of operations in replicating large numbers of identical units. I believe that the potential for cost reduction along this new avenue is substantial.

Five years later, the *EPRI Journal* contemplated "New Capacity in Smaller Packages" (732), mainly for reasons of financial risk management. Many of its member utilities were awaking with a bad financial hangover from the combination of nuclear binge, runaway capital-cost escalation, high inflation and interest rates (amidst aftershocks of the 1979 disruption, the prime rate averaged 18.9%/y in 1981), flagging demand growth, and soaring overcapacity. The industry's flagship research journal focused less on the engineering advantages of appropriate scale than on financial risk management, noting that "changing conditions are now prompting many utilities to take a fresh look at the matter of generating-unit size"—as if giant units were still preferable, just too risky. In particular, it noted,

> Uncertain load growth, constricted cash flow, and long lead times for large units define a new operating climate. It is risky to commit scarce capital to build a large unit that must be started many years in advance of the anticipated need....Today's financial climate requires a sharp match between capacity and demand because a major mismatch in either direction carries substantial cost. Building system capacity in small steps may be one way to optimize that match—hence, the growing interest among utilities in the concept of modular generation.

Improved system reliability (because many smaller—say, 100-MW—units were unlikely to fail simultaneously) and easier siting were also mentioned, though Fisher's inverse correlation between unit size and availability was not. EPRI's Dwain Spencer opined that:

> The concept of modular, parallel systems became a requirement and then a reality in order to achieve the high reliability required for missile and space missions. Now we have to demonstrate that this same idea can be applied to advanced power systems.

EPRI's Fritz Kalhammer saw "a broad trend toward integration of relatively small-scale, dispersed electricity sources into utility systems," and his colleague Kurt Yeager added that this trend looked durable over the long term, not a mere artifact of spiking interest rates.

Yet reflecting the ambivalence common in 1983, the article's author strongly emphasized coal combustion and coal gasification, even to run fuel cells (natural gas was then believed to be scarce and expensive). She thought the future of windpower—whose economic viability, she felt, remained to be established over the next five years—lay in gigantic 5-MW machines, which later turned out to exhibit strong technological diseconomies of scale.[2] She hoped phosphoric-acid fuel cells (the most advanced kind then contemplated—PEMFCs weren't mentioned) might "operate economically in increments as small as 10 MW"; their actual commercial scale today is 0.2 MW and falling. And she concluded, with a seeming wistfulness for the good old days, "Bigger will still be better in many applications, but as long as tight money and doubtful demand prevail, small modular units may fill a special need in prudent utility planning." As with Fisher, the overwhelming majority of the scale effects now known never got mentioned in that 1983 article; but piece by piece, the right questions were starting to be asked, even if "modular" often meant around 100–200 MW rather than much smaller.

All these themes, and many more, will emerge in the following discussion. But now, a quarter-century after John Fisher's regression analysis questioned the bigger-is-better dogma, diseconomies of scale are no longer mere tentative observations but a leading motivator of gigantic flux in the world's largest industry. Avoiding those diseconomies is increasingly emerging as a fount of quantifiable benefits that can reverse the merit order of economic choices. And making resources the right size, even if that's orders of magnitude smaller than tradition dictated, is emerging as the cornerstone of sound and profitable investments.

We begin with issues related to lead time—how long it takes to plan, site, get permits, and construct a power plant. To introduce that rich topic, we first survey the sources of uncertainty in electrical supply and demand on various timescales.

2.2.1 Many timescales, many uncertainties

The supply of electricity must be planned on a variety of timescales, ranging from a fraction of a second to decades. The reasons for this are physical, fundamental, and largely unavoidable.

Electricity is so difficult and expensive to store that except for a few special and costly large-scale installations, mostly using pumped hydroelectric storage, its supply is a real-time business (though that may change in this decade with new onsite technologies such as superflywheels and ultracapacitors (340) and even reversible fuel cells). In this respect, electricity differs from almost every other commodity. In effect, electricity is infinitely perishable—like bananas that must be eaten the very instant they are plucked, and ripened for plucking in exact coordination with the eaters' appetites. This inherent lack of inventory requires an understanding of all the diverse timescales on which those appetites may vary. We introduce this topic here in lay terms, then return to it more technically in Section 2.2.11.1 and Section 2.3.3.5 when discussing system stability and ramp rates. If you're not familiar with the operational fluctuations that electric power systems experience on a timescale ranging from milliseconds to days, please read Tutorial 1 now.

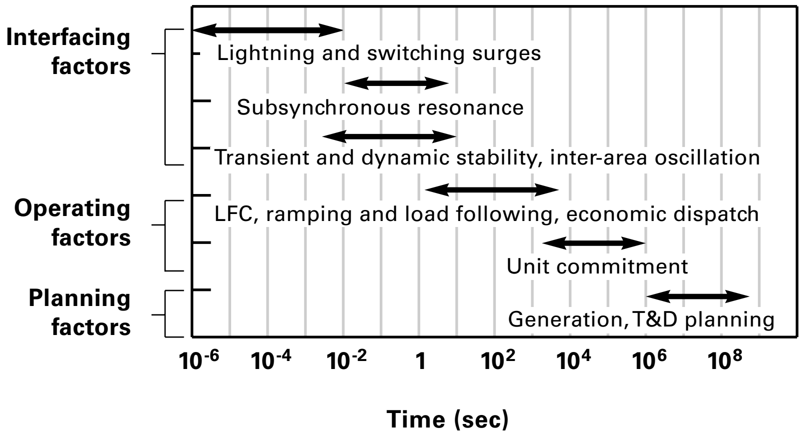
[2] By 1996 (688), commercial machines were typically rated at a few hundred kW; the largest commercial 1997 machines were 750 kW; and 1-MW machines were expected in prototype around 1998. They have since demonstrated some successes, but with caution and careful design. Earlier government-funded 2.5-MW machines, with near-supersonic tipspeeds and blades the size of jumbo-jet wings, were costly failures. Mid-1990s German engineering analyses (688) were finding cost minima around 30–40, or at most 60, meters rotor diameter, respectively corresponding to about 0.3–0.5, or at most ~1.3, MW; so on 2002 understanding of design and materials, 5-MW machines still look somewhat implausible.

Tutorial 1: Operational Fluctuations

Short-term fluctuations

Demand for electricity fluctuates from instant to instant as a myriad of users and controls unpredictably turn loads on and off. Supply may also fluctuate instantaneously as system faults, such as voltage spikes and interruptions caused by lightning or by sudden equipment failure, "shock" the grid. That shock then reverberates over distances ranging from local to vast, much like the wiggles in an enormous coupled system of weights connected by springs. Most of these fluctuations are offset by others fairly nearby, or occur on such a short timescale that they are smoothed out imperceptibly by the energy stored in the capacitance and inductance of the supply system.[3] They are the shortest of the timescales, down to microseconds, shown in Figure 2-1's graphic summary (699) of the timeframes relevant to power system management.

Figure 2-1: Electricity's timescales span 15 orders of magnitude
The timescales important to the planners and operators of electric supply systems span from microseconds to decades.



Source: Y. Wan and B. K. Parsons, "Factors Relevant to Utility Integration of Intermittent Renewable Technologies" (NREL, August 1993), p. 3

Longer timescales, on the order of one cycle or one "Hertz" (Hz)—in North America, 1/60th of a second or 17 milliseconds—traditionally require a specific and deliberate compensatory adjustment in supply or demand. Nowadays, transient stability on the transmission system, where even momentary glitches can cause vast quantities of power to slosh destructively around, is also requiring the evolution of new families of electronic power-switching and control devices. These can extend the same control and damping capability to a timescale of milliseconds, so that the grid can eventually act much like a giant integrated circuit—about a billion times bigger than conventional chips (328). This helps to deal with not only transient instability (the voltage oscillations caused by faults) and steady-state instability (overwhelming damping forces by transferring too much power through part of a transmission system), but also small-signal or dynamic instability. That's when normally unimportant variations in generation or load, too small to be considered disturbances, nonetheless trigger low-frequency oscillations that can grow into voltage and frequency fluctuations large enough to spoil system stability.

On the timescale of about a second or more, uncompensated changes in demand cause changes in the speed of rotation of the large turbo-alternators at steam or hydroelectric power stations: heavier demand takes angular momentum out of the rotors, causing them to slow down, while lighter demand unburdens them so they speed up. But the frequency of the alternating-current grid, which varies directly with the speed of the rotors, must be closely controlled in order to keep different generating units synchronized (with the "top" of each rotor reaching the straight-up position at the same instant as all the others) so they are all "pulling

[3] Chiefly the magnetic fields of transformers and conductors, and the energy storage of capacitors located mainly at the substations.

together": otherwise they could fight each other. If not immediately disconnected ("tripped offline") by protective relays, they could suffer disastrous loss of synchrony, cascading instability, and serious equipment damage.[4]

To maintain all the rotors within an acceptable "angular shift" (difference in instantaneous shaft angle) when a given rotor starts to slow down, its operator must in the short term adjust the excitation voltage to the rotor, and in the longer term promptly adjust the flow of steam or water or (in the case of gas turbines) fuel into the turbine to restore the normal operating speed before it departs from permissible limits.[5] (Gas turbines, being aerodynamic devices, can also stall if the shaft rotation slows down too much.) In practice, this is done by automatic generation control (AGC) coordinated by a vast telecommunications network that links devices at many different levels and locations, coordinating actions on a scale of milliseconds based on sensors whose data, in modern digital versions, are sampled up to 5,000 times per second (328). Conversely, if electrical demand decreases, the operator must correspondingly decrease the mechanical force driving the rotors, both to keep the frequency constant and to prevent them from spinning too fast (and, if that "overspeed" went uncontrolled, ultimately breaking apart—a risk if the unit isn't shut down within a fraction of a second of complete loss of its bus load [281]). The frequency must also be maintained at an average of exactly 60 Hz[6] over each 24-hour period; otherwise motor-driven electric clocks and other devices whose speed depends on grid frequency would gain or lose time. To keep this frequency rather exact, Load Frequency Control (LFC) checks and adjusts each governor's shaft speed every few seconds.

Grids currently handle these adjustments in the short term (up to a minute or so) by individual generators' shaft-speed controls, which operate automatically on a timescale of milliseconds, and by the centralized dispatch of **spinning reserve**—rotating and synchronized but not electrically loaded capacity specifically kept aside for this purpose. Additional **operating reserves** available by increasing the output of plants already operating and loaded, but not fully loaded, can also be brought online in periods ranging up to ten minutes, but often much less, since these resources are typically hydroelectric plants (which require valve-opening and rotor-spinup but no thermal warmup) and certain fast-start kinds of combustion turbines. Normally at least half of the total operating reserve is spinning, and the total operating reserve is adequate to cover the loss of the largest generating unit.

A "stability market" concept emerging first in New Zealand (303) adds a new way to meet such short-term operating requirements. Immediately interruptible loads, such as turning off an electric-resistance water heater on six seconds' notice, can be used to express the market value of offsetting other short-term increases in load, thereby stabilizing aggregate demand at significantly lower cost than could be done on the supply side (144, 399). That value is normally set by the cost of loading the spinning reserve. When the value is expressed in a two-way market, many interesting examples

[4] To ensure this, utility generators are almost always "synchronous" machines whose rotor current or "excitation" comes from a separate DC source or from the generator itself; with careful control, this explicit frequency control can keep all the rotors synchronized. In contrast, the induction generators used in some small-hydro and wind generators, and in many engine-driven generators, excite their rotors from an external AC source, typically the grid itself, thereby consuming reactive current (§ 2.3.2.3) so that they cannot generate without the grid's being energized.

[5] Those limits are a matter of convention, ranging from variations of less than 1 Hz to much larger values. Decades ago, frequency and phase stability limits were often said to be about an order of magnitude more stringent in North America than in Western Europe, where in turn they were about an order of magnitude more stringent than in Eastern Europe and the then Soviet Union. The lights stayed on (more or less) in all three regions across this wide range of operating philosophies: each simply dealt with the need for synchronization in different ways. In hindsight, it is not clear whether the more stringent control requirements in North American grids actually represented an economic optimum or only an unexamined assumption.

[6] The North American standard, although most of the rest of the world uses 50 Hz (50 cycles per second). Each cycle consists of a complete back-and-forth reversal of the alternating-current (AC) electric voltage "pressure" and the corresponding current flow.

of automated demand-side controls responding to real-time price signals start to emerge (515). These demand-side responses, the simplest of which are loads interruptible by underfrequency trips or by special signals, will become increasingly important and valuable in an electricity industry *dominated by its loads rather than by its generators*—a key characteristic that is already true today but not yet very widely recognized (303).[7]

On a slightly slower timescale than adjusting the steam or water valves, the power-plant operator or control system must adjust the fuel feed or combustion air, the nuclear reaction rate, or the dam's water flow. In the case of a steam plant, the steam temperature and pressure depend on the rate of combustion or nuclear reaction, requiring precise control of many interactive variables. In essence, however, all these controls are a fancy version of the old steam-locomotive boiler stoker who would shovel in coal more quickly to climb hills than to traverse level tracks. Power-station boilers, being very large metal objects, store heat and therefore have a thermal time constant that makes them respond only at a certain rate and with some delay that must be anticipated. Thermal power plants also use a large number of pumps, fans, and other devices that can change speed only with certain mechanical delays and changes in efficiency, becoming less efficient as they depart from the ideal operating conditions for which they were designed. The resulting control optimization is quite complex—especially in the case of a nuclear plant, where, for example, the nuclear reaction creates certain neutron-absorbing fission products that later inhibit the chain reaction until they gradually decay.

Complexities mount. In addition to the ramping up and down of various units to meet or anticipate loads while maintaining constant frequency, AGC also works on a longer timescale, typically 2–10 minutes, to adjust each generator's output to optimize the system's entire generating mix against various units' thermal efficiency, fuel and operating costs, and associated transmission losses, so that the incremental production cost of each generator in different parts of the system is equal (it is then called the system lambda). And in a rolling planning process called Unit Commitment, these considerations are integrated with longer-term requirements for scheduling the various generators to allow optimal maintenance, startup and shutdown costs, and minimum fuel-burn requirements to be met at lowest overall system cost. These criteria are typically reviewed daily and executed hourly, having regard to such longer-term considerations as seasonal availability and water storage in hydroelectric systems. But let us return to the shorter term.

Medium-term fluctuations

If a rising "ramp" of electrical demand cannot be satisfied simply by raising more steam in the plants already online, then the operator must start up additional generating capacity. In general, it takes much longer to start steam plants (like starting up a gigantic stove to get the water-kettle boiling) than to start engines or combustion turbines, so this non-operating reserve is traditionally defined as resources taking more than ten minutes to dispatch. Both for this reason and because of differing ratios of capital to operating costs, the operator typically has at her disposal a portfolio of different kinds of generating units. Based on her experience, she can "commit" (plan to start up) additional generating units in good time to meet required **ramp rates** (speed of increasing power output over time) at times of rising demand. Demand normally rises, for example, when people get to work in the mornings or come home and turn on appliances

[7] According to this compelling and important analysis, in future grid evolution, generators may be allowed to dispatch their output only if they provide, typically through a third-party aggregator of demand- and supply-side resources, an accompanying stability portfolio whose value is unbundled from the energy value. Otherwise they may be tempted to sell their spinning reserve margin into the profitable energy spot market rather than properly holding it back for the stability benefit of the system, and conversely, generators that provide vital stability services will not get properly compensated (303).

in the late afternoons, or when unusually hot or cold weather cause many electric heating or cooling systems to turn on more or less at the same time. Because very steep ramps may outrun the startup capabilities of the plant portfolio, utilities would be at risk of grid collapse if demand changed too quickly.[8]

This, then, is one aspect of the ever-changing operational task that utilities, running plants enormously larger than typical customers' loads, face throughout every day and night. But it is just the start of their wider planning challenge. They must carefully watch weather forecasts to ensure that, so far as possible, needed capacity will be available when severe weather causes peak system loads, rather than down for scheduled maintenance or at special risk of grid interruption by storms. Dispatchers must plan the weekly and seasonal variations of loads—adjusted for weather, strikes, holidays, major sporting events, even flu epidemics—to coordinate with fuel deliveries and inventories, maintenance, and other factors.

And then there is system planning for supply/demand balance over the long term—a big topic to which we turn next.

[8] For this reason, when a BBC producer in the 1970s wanted to invite viewers to go turn something off and observe the collective effect of these actions as displayed on a real-time meter of demand from the National Grid, the Central Electricity Generating Board successfully implored the BBC not to do so; it was already quite challenging enough for the grid's dispatchers to cope with the fast demand ramp that routinely occurred at the end of popular evening shows when millions of Britons would simultaneously get up from watching TV and go turn on their electric kettles to make a nice cup of tea.

2.2.1.1 Long-term supply/demand balances

Amidst the "noise" of short- and medium-term fluctuations in each kind of demand from each customer on many simultaneous timescales and with fine-grained geography, utility planners must also deal with secular trends. Changes in human populations with changing ages, household structures, needs, wishes, cultures, and end-use technologies all tend to change those people's amount and time patterns of electrical consumption.

Meanwhile, similar shifts occur on the supply side. Each year, some power stations may routinely reach the end of their useful lives, when they cost more to keep running than they are worth—though that balance is an ever-shifting function of technology, market conditions, and tax and regulatory policy. Some plants, too, may change their rated capacity: upwards ("repowering") with better control technologies, better boiler- or condenser-water chemistry, or higher-quality fuels, for example, or downwards ("derating") with corrosion, warmer

Benefits

1 *Distributed resources' generally shorter construction period leaves less time for reality to diverge from expectations, thus reducing the* probability *and hence the financial risk of under- or overbuilding.*

2 *Distributed resources' smaller unit size also reduces the* consequences *of such divergence and hence reduces its financial risk.*

3 *The frequent* correlation *between distributed resources' shorter lead time and smaller unit size can create a multiplicative, not merely an additive, risk reduction.*

condenser water caused by nearby heat sources or changing climate, fouling of heat-exchange surfaces, pollution restrictions, changes in nuclear safety rules, etc. And all kinds of surprises, from local to global, may dramatically alter the portfolio of plants and fuels available for use, on notice ranging from long to little to none.

This is no simple matter. Over the very prolonged timescale—traditionally a decade or more—for building a major new power station, it becomes more like what the military calls a SWAG (scientific wild-assed guess). Despite the most sophisticated forecasting methods, few if any electric utilities in the world have a consistently accurate record. Utility planners are not amused by physicist Niels Bohr's remark that "It is difficult to make predictions—especially about the future": in this business, major planning errors can compound to multi-billion-dollar mistakes from which an especially unfortunate utility might never recover. Having many *other* utilities (let alone non-utility producers) simultaneously making similar, but not necessarily coordinated, forecasts and investments to supply the same interconnected grid does not protect against each utility's own forecasting errors, and may make them worse by reinforcing a "herd instinct."

Here, however, an obvious benefit of distributed resources reveals itself. In general, smaller resources can be planned and built more quickly than very large ones; and the longer it takes to plan, site, and build a power station, the more likely reality is to diverge from forecasts (and on the larger scale corresponding to the size of the station itself), so the greater the likelihood and scale of under- or overbuilding, so the greater the financial risk of guessing wrong. That is (115),

> Inability to forecast precisely when power is needed involves a cost which is a function of the size *and* lead time of the units being considered and the relative flexibility provided by other units [or other resources such as demand-side management (DSM)] which the system can call on to bridge demand/supply gaps. Other things being equal, the larger the units, *and* the longer the construction lead times, the greater this cost will be, because it becomes more difficult to synchronize new power generating capacity with the growth in demand [over a larger increment and during a longer period].

Conversely, *the more closely the resource approaches the ideal of "build-as-you-need, pay-as-you-go," the lower the financial risk.*

It is important to note that this risk—of building too much or too little capacity to match demand—depends on unit size *and* on unit lead time. At least for conventional generating plants, these two variables are usually rather well correlated, so their risk-increasing effect is in principle multiplicative (though nonlinearly: only if lead time were uniformly proportional to unit size would risk rise exactly as the square of unit size). It might at first appear that the same is not true in reverse: smaller units tend to be faster (§ 1.5.7)—for much smaller distributed resources, very much faster—but they also can meet less demand, so to the extent their size and lead time are correlated (also nonlinearly), their risk-reducing advantage would be *reduced*. But this does not actually occur because small units are typically installed not singly but rather in large numbers that can *collectively* match (or more if desired) the "lumpy" capacity of the single large unit they displace. Therefore, in general, small units' risk-reducing effect is at least proportionate to their reduction in lead time, and will be even greater to the extent that large resources also take longer to build.

Chapman and Ward (115) correctly note that power planning takes place within "three separate planning horizons and processes"[9] that are "interdependent but separable, in the sense that they be considered one at a time in an iterative process, with earlier analysis in one informing the others." These three timescales, conceptually somewhat related to the scales of fluctuation described in Section 2.2.1 above, could be restated as:

- the short-term *operational* scale of keeping the grid stable, supply and deliverability robust, and the lights on, ranging from real-time dispatch to annual maintenance scheduling;
- the medium-term *planning* scale of keeping supply and demand in balance over the years through a flexible strategy of resource acquisition, conversion, movement, trading, renovation, and retirement; and
- the long-term *visionary* scale of ensuring over decades that the mix, scale, and management of energy systems are avoiding fundamental strategic errors; opening new options through farsighted RD&D and education; fostering a healthy evolutionary direction for institutional, market, and cultural structures, patterns, and rules; and sustaining foresight capabilities that will support graceful adaptation to and leadership in the unfolding future.

All three timescales are vital. So is not mixing them up. And so is seeking opportunities to serve synergistically the goals of more than one at a time, rather than creating tradeoffs between them. We therefore turn now to ways to value some specific attributes—modularity, modest scale, and short lead planning and installation times—of distributed resources that also happen to offer advantages on all three timescales and levels of responsibility.

2.2.2 Valuing modularity and short lead times

To reduce the financial risks of long-lead-time centralized resources, it is logistically feasible (§ 1.5.7) to add modular, short-lead-time distributed resources that add up to significant new capacity. But can those smaller resources create important economic benefits by virtue of being faster to plan and build? Common sense says yes, and suggests three main kinds of benefits: reducing the *forecasting risk* caused by the unavoidable uncertainty of future demand; reducing the *financial risk* caused directly by larger installations' longer construction periods; and reducing the *risk of technological or regulatory obsolescence*. Let us consider these in turn.

2.2.2.1 Forecasting risk

Nearly twenty years ago, M.F. Cantley noted that "The greater time lags required in planning [and building] giant power plants mean that forecasts [of demand for them] have to be made further ahead, with correspondingly greater uncertainty; therefore the level of spare capacity to be installed to achieve a specified level of security of supply must also increase." (90) Longer lead time actually incurs a double penalty: it increases the uncertainty of demand forecasts by having to look further ahead, *and* it increases the penalty per unit of uncertainty

[9] They add that "Additional (four or more) horizons might be usefully explored, but fewer than three will cause difficulties."

Benefits

4 *Shorter lead time further reduces forecasting errors and associated financial risks by reducing errors' amplification with the passage of time.*

5 *Even if short-lead-time units have lower thermal efficiency, their lower capital and interest costs can often offset the excess carrying charges on idle centralized capacity whose better thermal efficiency is more than offset by high capital cost.*

6 *Smaller, faster modules can be built on a "pay-as-you-go" basis with less financial strain, reducing the builder's financial risk and hence cost of capital.*

7 *Centralized capacity additions overshoot demand (absent gross underforecasting or exactly predictable step-function increments of demand) because their inherent "lumpiness" leaves substantial increments of capacity idle until demand can "grow into it." In contrast, smaller units can more exactly match gradual changes in demand without building unnecessary slack capacity ("build-as-you-need"), so their capacity additions are employed incrementally and immediately.*

8 *Smaller, more modular capacity not only ties up less idle capital (#7), but also does so for a shorter time (because the demand can "grow into" the added capacity sooner), thus reducing the cost of capital per unit of revenue.*

9 *If distributed resources are becoming cheaper with time, as most are, their small units and short lead times permit those cost reductions to be almost fully captured. This is the inverse of #8: revenue increases there, and cost reductions here, are captured incrementally and immediately by following the demand or cost curves nearly exactly.*

10 *Using short-lead-time plants reduces the risk of a "death spiral" of rising tariffs and stagnating demand.*

by making potential forecasting errors larger and more consequential. As *Business Week* put it in 1980 (83), "Utilities are becoming wary of projects with long lead times; by the time the plant is finished, demand could be much lower than expected. If you're wrong with a big one, you're really wrong.... Uncertainty over demand is the main reason for the appeal of small plants."

This forecasting risk became painfully evident in the 1970s, when the power industry consistently overestimated demand growth while lead times for large new generating plants became longer and more uncertain, the cost of capital soared, and utilities used planning models "biased toward large plants." The interaction of these four factors created "an increased likelihood of excess capacity, unrecoverable costs and investment risk" (373) that bankrupted a few utilities and severely strained scores more. The industry therefore learned the hard way that minimizing risk "will tend to favor smaller scale projects, with shorter lead times and less exposure to economic and financial risks." (373) Specifically (373):

- An autumn 1978 *Energy Daily* review (522) of data collected by the Edison Electric Institute in autumn 1978 showed that only once in the previous 11 years had the industry underpredicted the following year's total noncoincident peak demand, and then only by 0.1 percentage point. Rather, the forecasts averaged 2.1 percentage points too high during 1968–73 and 5.1 percentage points too high after 1974. Indeed, during 1974–79, the average forecast error exceeded the average annual growth rate, and during 1975–78 the error averaged 2.5 times the actual growth—leading the editor of *Electrical World* to call for a major rethinking of traditional forecasting methods (289) (see Figure 1-41 in Part One).

- In such an uncertain forecasting environment, "The alternative to waiting 12 years to see whether demand growth did justify construction of an expensive large generator...is building smaller projects with shorter lead times." (522) For example, if a utility forecast 5.5% annual demand growth, built new generators with 12-year lead times, and actually experienced only 3.5% annual demand growth, then it would end up with 26% excess capacity. If the lead time were 6 years, however, that excess would drop to 12%; if 4 years, to 8%.
- Lead time correlated well with unit size: *e.g.*, for U.S. coal-fired plants in the 300–700-MWe range, each 100 MW of capacity required an extra year of construction. Although different analysts' values for this coefficient vary,[10] the existence of an important bigger-hence-slower correlation has long been well established (12, 557).

For these reasons, as summarized by Sutherland *et al.* (673), with emphasis added,

> The most important result is that short lead time technologies, which represent smaller units, are a defense against the serious consequences of unforeseen changes in demand. The "worst case" occurs when electric utilities build large and long lead time plants [but]...anticipated demand is unrealized. A price penalty is paid by consumers, and unfavorable financial conditions plague the utility. Ford and Yabroff (1980, 78) concluded that the strategy of building small, short lead time plants could cut the price penalty to the consumer by 70% to 75%. *Both demand uncertainty and short lead times favor small generating units, with their synergistic effects being the most important.*

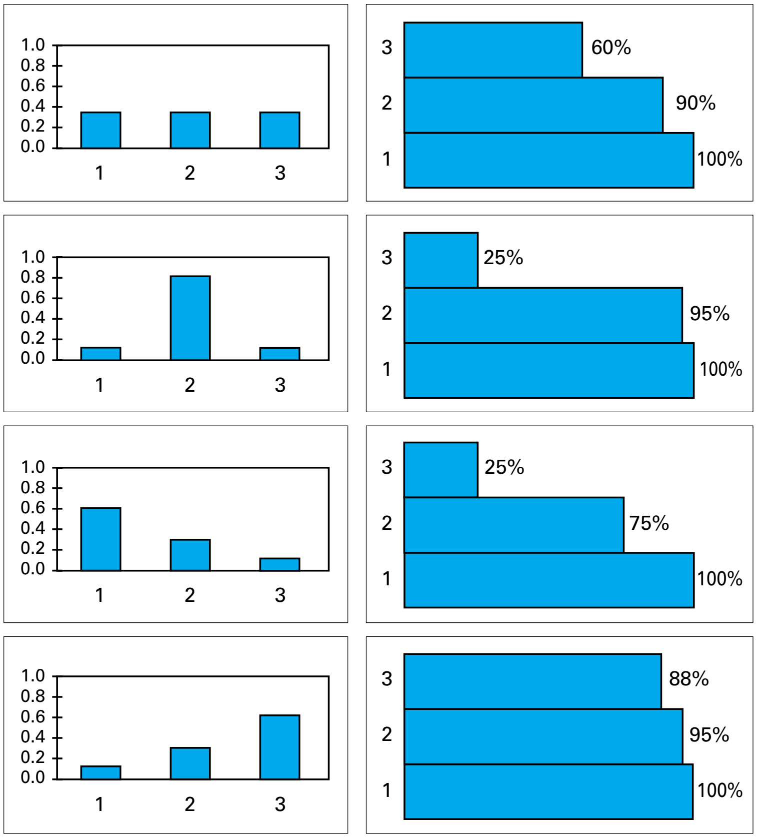
The mechanisms of that synergy become more visible when one looks more closely into the details of demand uncertainty. A lucid analysis of the tradeoffs between hoped-for power-plant economies of scale and the risk of excess capacity (75) (Figure 2-2) provides cost ratios showing how much cheaper the output from a larger unit must be, if it takes twice as long to build as a small plant, in order to justify buying the large plant under a given pattern of demand uncertainty. That pattern is expressed as the probability that during the planning period, demand will grow by one, two, or three arbitrary units, which can be interpreted as relative percentage growth rates. Those probabilities can occur in various combinations. For each, a set of ratios shows how much cheaper the large plant must be than the small plant in order to justify building the large one. In general, the assumed demand growth will justify at least one large unit. But to justify a second or third large unit, it must be modestly or dramatically cheaper than the smaller units, depending on the distribution of demand probabilities. The left-hand graph in each case shows the assumed distribution of probabilities (for example, in the first case, all three demand growth rates—*e.g.*, *x*, 2*x*, and 3*x*—are equally probable). The right-hand graph shows in the first case,

[10] For example (673), a RAND multiple-regression analysis by William Mooz found a correlation equivalent to ~3.5 months of construction duration per 100 MWe of net capacity (but actually a bit nonlinear), while a comparable analysis in a different algebraic form, by Charles Komanoff, found that a doubling of nuclear unit size would increase construction time by 28%. (Komanoff's capital-cost model for coal plants didn't use unit size as a variable, but unit size was the variable most significant in affecting construction duration.) A further analysis cited (673), using an EPRI database of 54 coal and nuclear plants, didn't examine unit size as an explanatory variable, but did find that 22% of the nuclear units' construction delay was deliberate in an effort not to build too far ahead of demand, implying that "the utility would have been better off with smaller and shorter lead time plants."

for instance, that a large unit is justifiable at full cost as the first unit to be built, but must be 10% cheaper than the small plant to be the right choice as the second unit, and 40% cheaper as the third unit.

Figure 2-2: Uncertain demand imposes stringent cost tests on slow-to-build resources

Long-lead-time power stations must be far cheaper than halved-lead-time smaller units in order to be an economical way to keep on meeting changing demand (unless, perhaps, demand growth is known to be accelerating).



Source: E. P. Kahn, "Project Lead Times and Demand Uncertainty: Implications for Financial Risk of Electric Utilities" (Lawrence Berkeley Laboratory/University of California, 1979), p. 9, fig. 4

Thus continuing to build large plants requires them to be built at an increasingly steep cost discount even if demand growth is steady (the first case); is unlikely to be the right strategy if demand fluctuates markedly (the second case) or demand growth tapers off (the third case); and may be justifiable if demand growth is definitely and unalterably accelerating (the fourth case). This comparison—focusing only on a specific kind of investment risk, and not taking account of several dozen other effects of scale on economics—is of course a simplified illustration of planning choices that could be simulated more elaborately, typically by a Monte Carlo computer analysis. But simple though it is, the example starkly illustrates the risks of overreliance on long-lead-time plants when demand is uncertain: in the middle two cases, the third large unit could be justified only if it were *fourfold cheaper* than the competing small, halved-lead-time unit. The authors conclude (75):

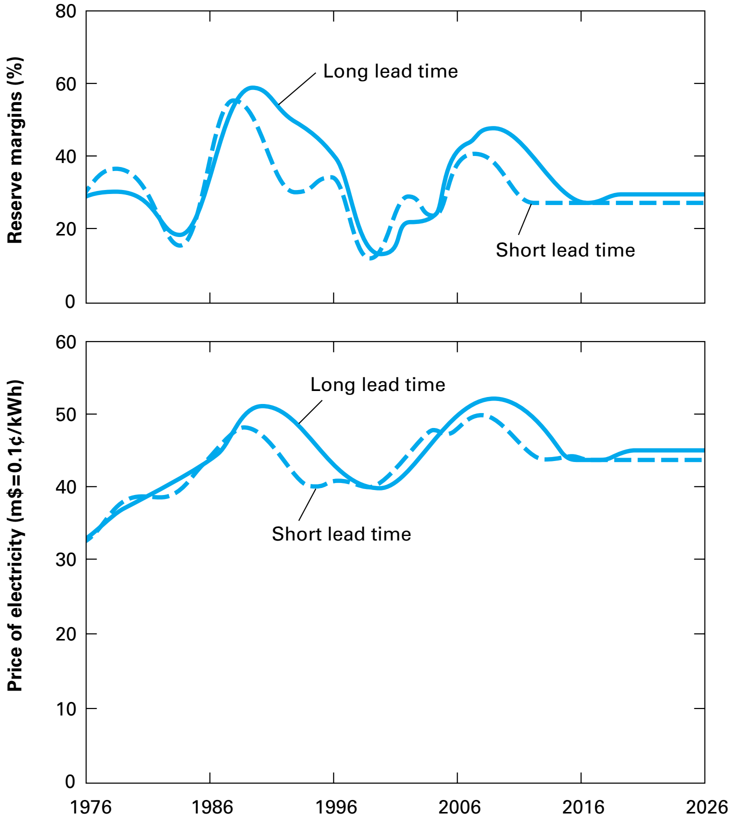
> The relative cost advantage of short lead time plants can be substantial. If demand uncertainty is such that low growth rates of demand are more likely than high growth rates, or if the variance in demand growth is simply large, the capital cost of long lead time plants must be substantially decreased, under some circumstances as much as 50%[,] to make long lead time plants cheaper, even with a flat load curve. The fraction of future demand that is optimally satisfied with long lead time power plants depends on two factors. Again, the lower the probability that a given level of demand will occur, the greater the cost advantage required to make long lead time plants optimal for that level. This conclusion is modified by the existing mix of short lead time—high [fuel] cost plants and long lead time—low fuel cost plants. The more short lead time plants in the existing mix[,] the smaller the cost advantage of long lead time plants needs to be. In general[,] unless long lead time plants have a substantial cost advantage or the probability of the demand['s] growing at the maximum rate is large, it is rarely optimal to supply all the projected demand with long lead time plants.

In summary: if too many large, long-lead-time units are built, they are likely to overshoot demand. Paying for that idle capacity will then raise electricity prices, further dampening demand growth or even absolute levels of demand, and increasing pressure for even further price increases to cover the revenue shortfall. This way lies financial crisis, as the industry found to its cost in the 1970s and 1980s.

Of course, forecasting errors go both ways: you can build capacity that you turn out not to need, or you can fail to build a plant that you *do* turn out to need. Are those risks symmetrical? In the 1970s, when power-plant (especially nuclear) vendors were trying to justify their seemingly risky GW-range products, they cited studies purporting to show that underbuilding incurred a greater financial penalty than overbuilding (100, 671). However, those studies' recommendation—to overbuild big thermal plants as a sort of "insurance" against uncertain demand—turned out to result from artifactual flaws in their models (243, 249, 417).[11] More sophisticated simulations, on the contrary, showed that (at least for utilities that don't start charging customers for power plants until they're all built and put into service) if demand is uncertain, financial risk will be minimized by deliberately *under*building large, long-lead-time plants (75, 243–4, 246–7, 249).

For example, given an illustratively irregular pattern of demand growth characteristic of normal fluctuations in weather and business conditions, excessive reserve margins and electricity prices can be reduced by preferring short-lead-time plants (Figure 2-3):

Figure 2-3: Faster-to-build resources help avoid capacity and price overshoot
Short-lead-time plants help to avoid excessive reserve margins and tariffs under uncertain demand.



Source: A. Ford and A. Youngblood, "Simulating the Planning Advantages of Shorter Lead Time Generating Technologies" (*Energy Systems and Policy* 6, 1982), p. 360, figs. 7 and 8

[11] The EPRI models assumed that all forms of generating capacity are expanded at the same rate, so that baseload shortages automatically incur [large] outage costs rather than extending the capacity or load factor of peaking or intermediate-load-factor plants. (This assumption means that the plant-mix questions at issue simply cannot be examined, because plants are treated as homogeneous.) Furthermore, the use of planning reserve margin as the key independent variable obscured the choice between plants of differing lead times. Capital costs were assumed to be low, so that even huge overcapacity didn't greatly increase fixed costs. Outage costs were treated as homogeneous, even though it would make more sense to market interruptible power to users with low outage costs. Uncertainties were assumed to be symmetrical with respect to under- or overprediction. And the opportunity costs of over- or underbuilding were ignored, whereas in fact, overbuilding ties up capital and hence foregoes the opportunity to invest in end-use efficiency or alternative supplies, while underbuilding means one still has the capital and can invest it in ways that will hedge the risk. For further comparative discussion of conflicting studies, see (249).

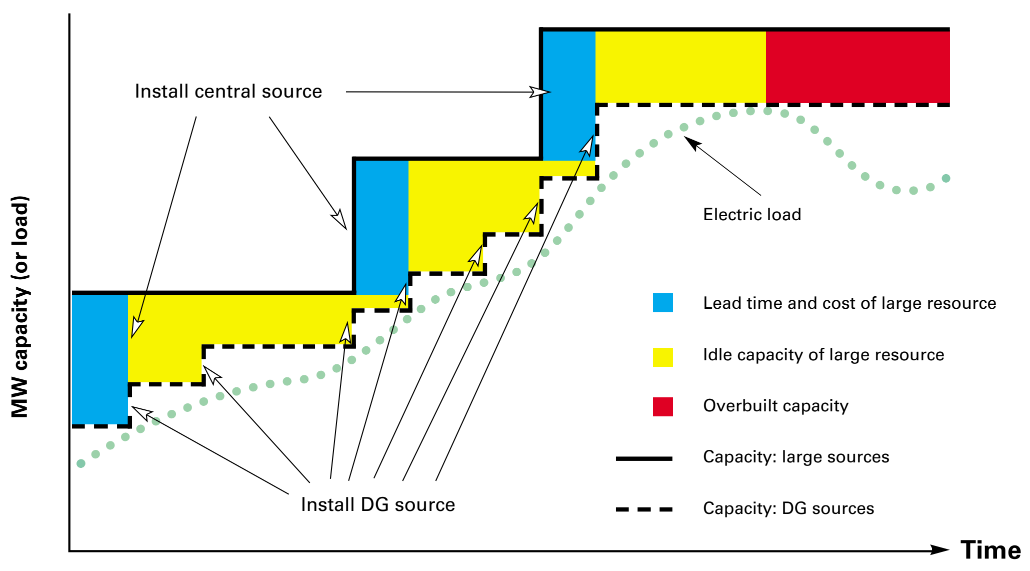
There are four reasons for this:

- operating short-lead-time, lower-thermal-efficiency, low-capital-cost stopgap plants (such as combustion turbines fueled with petroleum distillate or natural gas) more than expected, and paying their fuel-cost penalty, is cheaper than paying the carrying charges on giant, high-capital-cost power plants that are standing idle;[12]
- even if this means having to build new short-lead-time power stations such as combustion turbines, their shorter forecasting horizon greatly increases the certainty that they'll actually be needed, reducing the investment's "dry-hole" risk;
- smaller, faster modules will strain a utility's financial capacity far less (for example, adding one more unit to 100 similar small ones, rather than to two similar big ones, causes an incremental capitalization burden of 1%, not 33%); and
- short-lead-time plants can be built modularly in smaller blocks (301), matching need more exactly.

This last point is so obvious that it is often overlooked: big, "lumpy" capacity additions *invariably* overshoot demand (absent gross underforecasting of rapidly growing demand), leaving substantial amounts of the newly added capacity idle until demand can "grow into it" (Figure 2-4).[13]

Thus adding smaller modules saves three different kinds of costs: the increased lead time (and possibly increased total cost) of central resources; the cost of idle capacity that exceeds actual load; and overbuilt capacity that remains idle. Both curves maintain sufficient capacity to serve the erratically growing load, but the small-module strategy does so more exactly in both

Figure 2-4: Slow, lumpy capacity overshoots demand in three ways
The yellow areas show the extra capacity that big, lumpy units require to be installed before they can be used. Small distributed-generation (DG) modules don't overshoot much; they can be added more closely in step with demand. The blue areas show the extra construction and financing time required by the longer-lead-time central units.



Source: J. N Swisher, "Cleaner Energy, Greener Profits: Fuel Cells as Cost-Effective Distributed Energy Resources" (RMI, 2002), www.rmi.org/sitepages/pid171.php

[12] Naturally, this sort of conclusion is not immutable, but rather depends on interest rates, fuel costs, and other factors that change over time.

[13] This is quite an old and familiar problem in mathematical economics (588, 657). The latter paper concludes that "efficient production when there is uncertainty of demand forces the supplier to sacrifice economies of [unit] scale in order to achieve greater flexibility through a larger number of plants. Equally important is the result that full efficiency requires a set of plants of different sizes. Thus there is no optimal scale of plant or minimum efficient scale and in fact such a concept is meaningless in the present context. Only the collection of all plants is efficient."

quantity and timing, and hence incurs far lower cost.

This load-tracking ability has value unless demand growth not only is known in advance with complete certainty, but also occurs in step-functions exactly matching large capacity increments. If that is not the case—if the growth graph is diagonal rather than in vertical steps, even if it is completely smooth—then smaller, more modular capacity will tie up less idle capital for a shorter period.

If demand grows steadily, the value of avoiding lumps of temporarily unused capacity can be estimated by a simplified method modified by Hoff, Wenger, and Farmer (324) from a 1989 proposal by Ren Orans. The extra value of full capacity utilization is proportional to:

$$\frac{T(d-c)}{1-e^{-T(d-c)}}$$

where d is the [positive] real discount rate, c is the real rate at which capacity cost escalates, and T is years between investments. This approximation yielded reasonable agreement with PG&E's estimate (§ 2.3.2.6) for deferring Kerman transformer upgrades (324).

This analysis also provides a closed-form analytic solution for the case where the distributed resource is becoming cheaper with time, so even if it's not cost-effective now, it is expected to become so shortly. If the relative rates of cost change between the distributed and traditional resources are known, due allowance can be made. The equations provided (324) can also use option theory (§ 2.2.2.5) to account for uncertainties in the cost of the distributed resource. Such uncertainty may create additional advantage by suitably structuring the option so that the manager is entitled but not obliged to buy, depending on price. For these reasons, in an actual situation examined, a distributed resource costing $5,000/kW can be a cost-effective way to displace generating investments that would otherwise be made annually, plus transmission investments that would otherwise be made every 30 years—largely because the lumpiness of the latter investment means paying for much capacity that will stand idle for many years.[14]

In any actual planning situation, depending on the fluctuating pattern of demand growth, the extra cost of carrying the lumpy idle capacity can be calculated from the detailed assumptions, and then interpreted as a financial risk. Some tools for this calculation are described below. In principle, but not in most models, such a calculation should take into account an important economic feedback loop—the likelihood that the higher electricity tariffs needed to pay that extra cost will make demand growth both less buoyant and less certain, further heightening the financial risks (247–8). This sort of feedback is probably best captured by system dynamics models (248). Those models broadly confirm the "death spiral" scenario characteristic of plants that take longer to build than it takes customers to respond to early price signals from the costly construction—especially if demand is as sensitive to price as many econometric analyses suggest.[15] Avoiding the risk of the "death spiral" is an important potential benefit.

[14] It's important for the analytic tools used in this situation to capture declining costs incrementally and immediately, so that no cost reduction is delayed or lost through stepwise capture at longer intervals.

[15] Econometric studies collected by Ford and Youngblood (248) found long-run own-price elasticities of demand as large as –1.5 in the residential and commercial sectors and –2.5 in the industrial sector, with widely varying time constants. In general, elasticities with an absolute value larger than unity can lead to trouble; many of the values cited, including most of the industrial ones, are in this range. (An elasticity of –1.5 means that each 1% increase in price leads to a 1.5% decrease in demand. "Own-price" refers to the price of the same commodity whose demand is being measured; that differs from "cross-price" elasticities, which describe substitution of one resource for another as their relative prices change. "Long-run" typically refers to a period of years.)

Benefits

11 *Shorter lead time and smaller unit size both reduce the accumulation of interest during construction—an important benefit in both accounting and cashflow terms.*

12 *Where the multiplicative effect of faster-and-smaller units reduces financial risk (#3) and hence the cost of project capital, the correlated effects—of that cheaper capital, less of it (#11), and needing it over a shorter construction period (#11)—can be triply multiplicative. This can in turn improve the enterprise's financial performance, gaining it access to still cheaper capital. This is the opposite of the effect often observed with large-scale, long-lead-time projects, whose enhanced financial risks not only raise the cost of project capital but may cause general deterioration of the developer's financial indicators, raising its cost of capital and making it even less competitive.*

13 *For utilities that use such accrual accounting mechanisms as AFUDC (Allowance for Funds Used During Construction), shorter lead time's reduced absolute and fractional interest burden can improve the quality of earnings, hence investors' perceptions and willingness to invest.*

14 *Distributed resources' modularity increases the developer's financial freedom by tying up only enough working capital to complete one segment at a time.*

15 *Shorter lead time and smaller unit size both decrease construction's burden on the developer's cashflow, improving financial indicators and hence reducing the cost of capital.*

16 *Shorter-lead-time plants can also improve cashflow by starting to earn revenue sooner—through operational revenue-earning or regulatory rate-basing as soon as each module is built—rather than waiting for the entire total capacity to be completed.*

17 *The high velocity of capital (#16) may permit self-financing of subsequent units from early operating revenues.*

18 *Where external finance is required, early operation of an initial unit gives investors an early demonstration of the developer's capability, reducing the perceived risk of subsequent units and hence the cost of capital to build them.*

19 *Short lead time allows companies a longer "breathing spell" after the startup of each generating unit, so that they can better recover from the financial strain of construction.*

20 *Shorter lead time and smaller unit size may decrease the incentive, and the bargaining power, of some workers or unions whose critical skills may otherwise give them the leverage to demand extremely high wages or to stretch out construction still further on large, lumpy, long-lead-time projects that can yield no revenue until completed.*

21 *Smaller plants' lower local impacts may qualify them for regulatory exemptions or streamlined approvals processes, further reducing construction time and hence financing costs.*

22 *Where smaller plants' lower local impacts qualify them for regulatory exemptions or streamlined approvals processes, the risk of project failure and lost investment due to regulatory rejection or onerous condition decreases, so investors may demand a smaller risk premium.*

23 *Smaller plants have less obtrusive siting impacts, avoiding the risk of a vicious circle of public response that makes siting ever more difficult.*

(End of excerpt)