THE S-CURVE: A NEW FORECASTING TOOL

It is quite impossible that the noble organs of human speech could be replaced by ignoble, senseless metal.

—Jean Bouillaud, member of the French Academy of Science, regarding Thomas Edison's phonograph, 1978

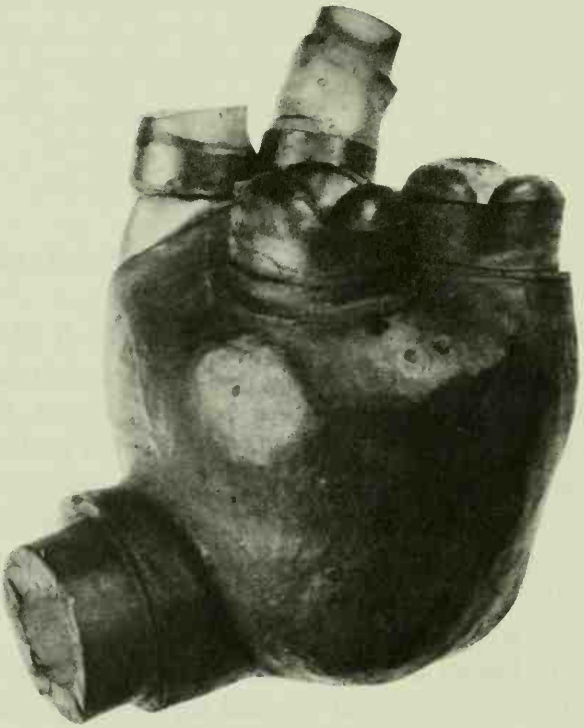
HEARTS AND MINDS

In December of 1982, at the University of Utah Medical Center in Salt Lake City, Dr. William DeVries replaced Dr. Barney Clark's diseased heart with an artificial one. Not a metal heart to be sure, but an artificial one that performed nobly none the less. In one sense it was the culmination of work begun many years earlier. But in another sense it is the story of the bottom end, or ascending portion, of the S-curve. It provides insight into how an S-curve looks and feels to those who are trying to push it ahead.

The story begins in 1957 in the Cleveland Clinic. The Cleveland Clinic is a referral hospital for special problems, especially those of the heart. It is a sprawling complex of a dozen buildings spread over 100 acres on the edges of downtown Cleveland. Patients have included President Sekou Touré of Guinea, King Hussein of Jordan, the King of Bhutan, the royal family of Nepal and the King of Saudi Arabia.

It was at the Cleveland Clinic that Willem Kolff and Tetsuzo Akutsu began their research with artificial hearts. In one of their first experiments, they used a plastic sack for a heart to keep a dog alive for 90 minutes (Exhibit 5).

They then began using larger animals, but the first artificial hearts they developed had two major problems. First, they were externally squeezed to pump the blood; the squeez-



5 Kolff and Akutsu Heart, 1957.

ing crushed most of the blood cells (which doctors call hemolysis) that passed through them. Those cells that survived saw the heart as an invader in the body and tried to defend the body by bringing white blood cells to the heart. The resulting coagulation stopped the blood flow. The animals all died quickly after the artificial hearts were implanted.

The doctors tried new design concepts to overcome these problems, but their attempts did not find quick success. As Dr. Robert Jarvik recalled in 1982:

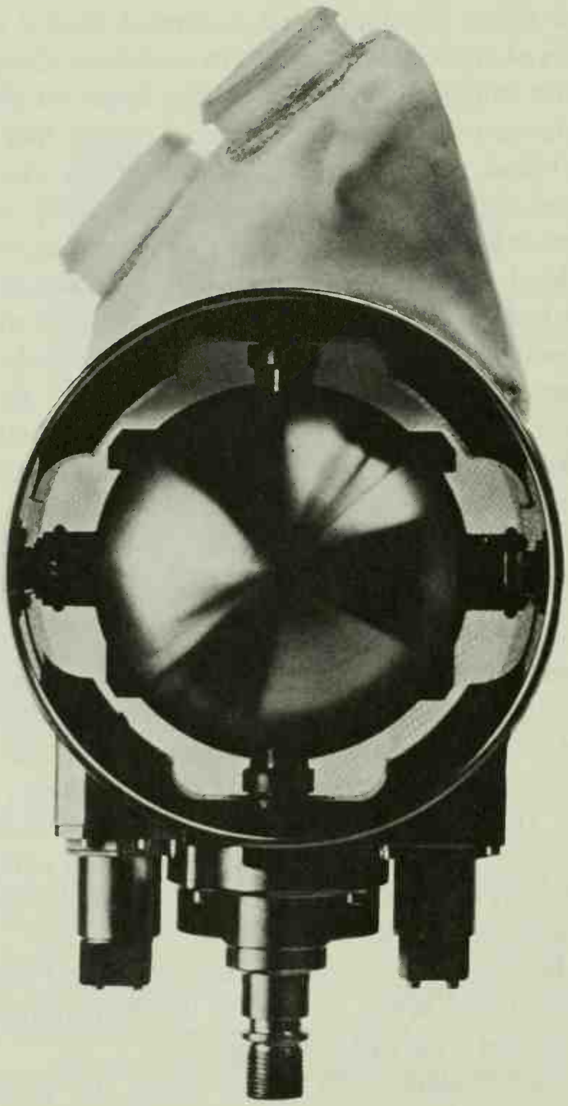
> Over the next few years Kolff's group at the Cleveland Clinic developed and implanted several other kinds of artificial heart driven by electricity. One such device employed five solenoids that displaced oil; the oil in turn compressed polyurethane sacks that held the blood. Animal survivals of three hours were obtained with this heart. In another electrically driven heart an electric motor drove a roller that compressed a blood-carrying tube against a foam-lined housing. The heart needed only outflow valves, but it caused excessive hemolysis and sustained life for only two hours. In the pendulum heart a pivoting electric motor alternately compressed two blood-containing sacks, thereby forcing blood out of the ventricles. Several dogs survived for from four to six hours with this device, but its output was inadequate and it caused excessive hemolysis.*

Other approaches were also tried. One used a nuclear-powered impeller that pumped the blood (Exhibit 6), but the impeller also crushed the cells. Another attempt had rubber sacs activated by compressed air (Exhibit 7). It too crushed the cells. Then, in 1970, Clifford Kwan-Gett, who had been with Kolff at the Cleveland Clinic and moved with him to the University of Utah in 1967, solved the problem by gently pumping the blood with a diaphragm (Exhibit 8).

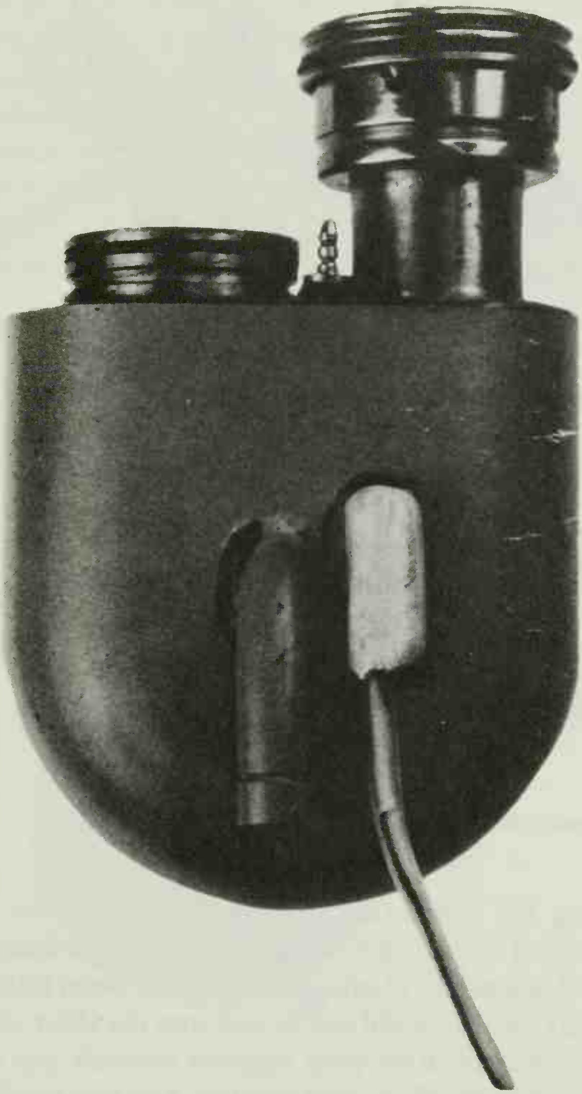
But the uncrushed blood cells still coagulated on the surface of the heart. While better than the Kolff-Akutsu heart, the Kwan-Gett wasn't good enough.

Another researcher, Dr. Nosé, solved this problem by

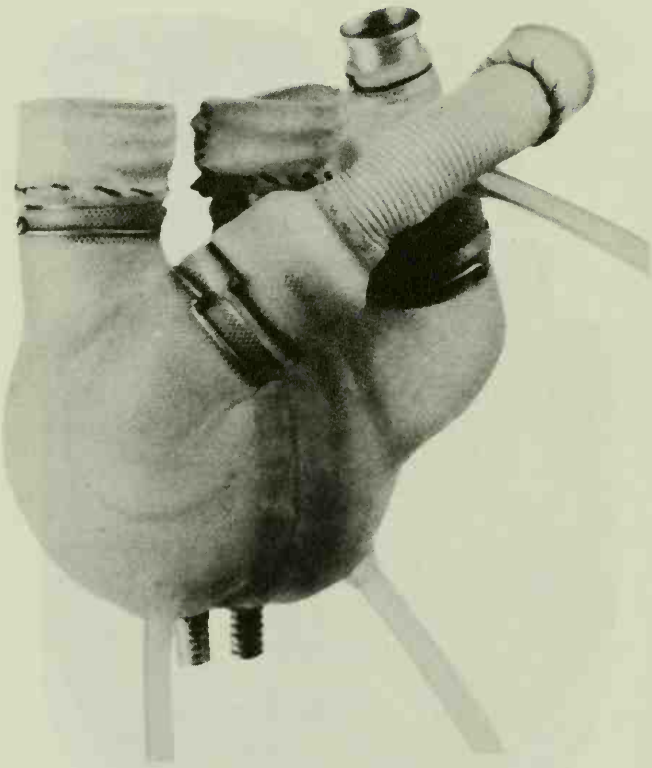
* Jarvik, R. K., "The Total Artificial Heart," *Scientific American*, January 1981, p. 77.



6 Artificial Heart.
Developed by Westinghouse and the University of Utah, 1975.



7 Yukihikio Nosé Heart, 1965.

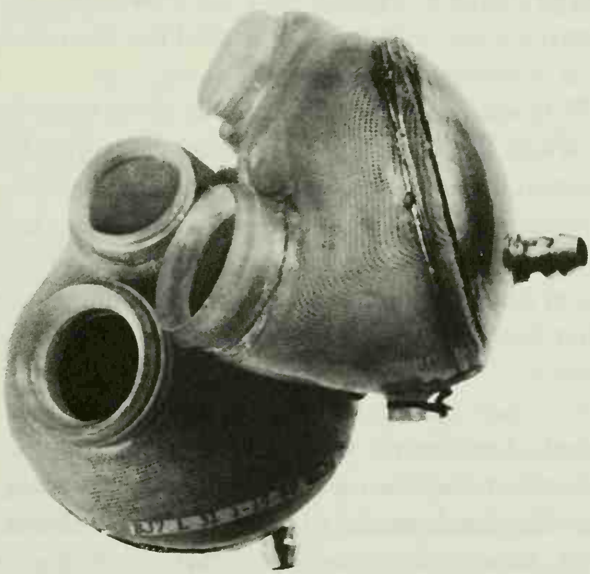


8 Kwan-Gett Heart, 1967.

replacing Dr. Kwan-Gett's synthetic silicon heart interior with natural tissues (such as the outer heart membrane). This approach solved the clotting problem, but blood leaked from this heart because it did not fit well into the chest cavity. As happens so often in the early stages of research, one removes one impediment only to find another. And that is why the S-curve is so flat in the beginning.

Dr. Jarvik, who as a high school student was improving methods to staple wounds shut, designed a heart that fit the chest cavity much better (Exhibit 9). The experimental animals now lived much longer than before. Real progress had been made. Jarvik began to ascend the S-curve of artificial heart technology.

Another problem struck. Now that the heart kept patients alive longer, the diaphragms began to give out. Jarvik replaced the rubber diaphragms with Lycra, an elastic material used in bras and girdles. It had the necessary wear resis-



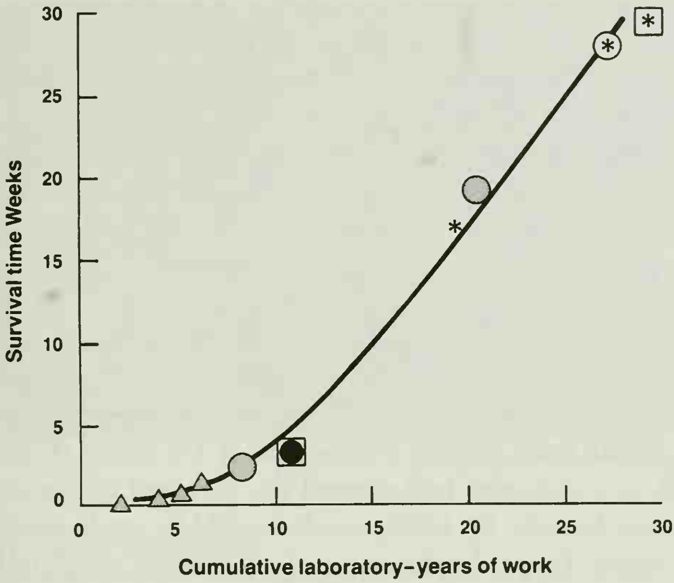
9 The Jarvik-7 Heart.

tance. The heart life increased up to four months and later to six months. Subsequently, Jarvik replaced the Lycra diaphragm with a specially designed rubber, and that lasted even longer.

With this evidence, Dr. Kolff and the University of Utah submitted an application for approval to implant an artificial heart in a human in February 1981. After an initial rejection by the FDA in March (for reasons that had nothing to do with the heart's performance but rather with questions about the psychological impact of the implanted heart on the patient), the application was approved in June. Then in 1981, Barney Clark had his heart replaced by DeVries. He lived for 112 days before he died of pneumonia. William Schroeder received the next artificial heart on November 25, 1984. He is surviving but is depressed, as the FDA feared he might be. But he is alive.

With each trial the performance of the heart became better. When the heart's performance, measured by the post-operation lifetime of the patient, is plotted versus the effort the various medical teams put into improving the hearts, the beginnings of an S-curve appear (Exhibit 10). With each subsequent heart representing a new data point, progress is slow at first but then accelerates rapidly as Jarvik makes his advances.

For each new product (or process) the S-curve shows precisely how much performance has improved and how much effort has been expended to gain that improvement. What the development of artificial hearts shows is that S-curves, though they are abstractions, can also be graphic histories of human efforts to solve problems. They represent the trials and errors of talented people. Displaying their efforts allows us to see patterns of success and failure that we



10 Artificial Heart Research.
The early stage of the S-curve for artificial hearts—it represents the inevitable trials and errors of research and invention.

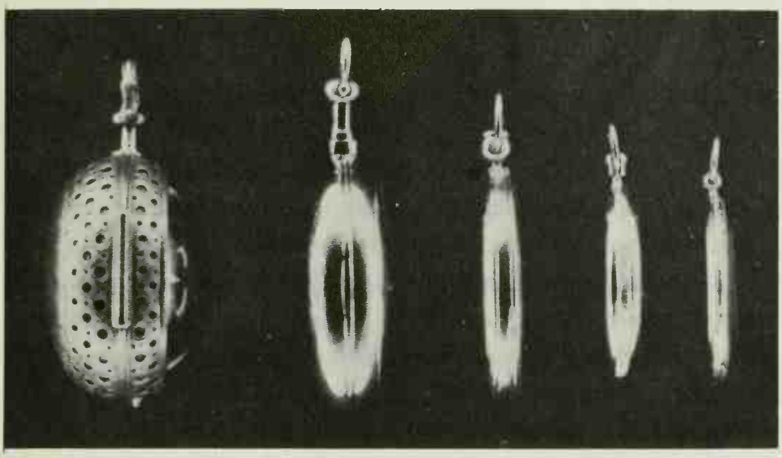
wouldn't otherwise find. In this case a rather long period of little progress followed by growing success.

DIMINISHING RETURNS

The process can also work the other way as technological limits are approached. Rather than showing more and more progress with less and less effort, each new step makes less and less progress. One of the most interesting cases to me of diminishing returns took place over a century ago and involved pocket watches. The "pocket" watch, which was itself something of a minor revolution (clocks were meant to be in towers, not in one's pocket), appeared about 1600. The first models were the shape and size of a lemon. For the gentleman carrying them, they provided the convenience of knowing time with precision, but they did create a rather unsightly bulge in his trousers. Accordingly, it became the fashion to design, or attempt to design, thinner watches. Put in our terms, thinness became the performance measurement. Watch designers put in effort to create ever thinner watches.

And they succeeded, but with diminishing returns, as Exhibit 11 shows. In 1700 when watch production, which is a surrogate term for "effort," was quite low, both the French and British were making watches about 1½ inches thick. By 1800, as production had stepped up, they had gotten down to about ¾ inch. By 1850, watches could be made about ¼ inch thick. That's the thickness of my "wrist" watch today! Not much different from 1850. If we take the nineteenth-century continental gentleman's definition of performance, thinness, its limit was reached in 1850 and before that came diminishing returns. Each new design represented efforts to pack more and more components into smaller and smaller spaces. As thinness reached its limits, other performance parameters (area, reliability, easy use, and cost) gained in importance.

These two stories, artificial hearts and pocket watches, when put together describe the two major phenomena behind the sinewy shape of the S-curve: learning (artificial hearts) and diminishing returns (pocket watches). The S-curve traces out the path of development of new products and processes with each successive point on the curve representing an improvement in performance. The pattern of the S-curve repeats itself again and again in industry after industry. These empirical observations, coupled with the underly-



c. 1700 1812 c. 1815 to 1825 1846 1850

11 British and Swiss Watches.
Pocket watches started reaching the limit to their thinness around 1850. Other performance parameters such as reliability, easy use and cost then gained in importance.
Source: Landes, David S., *Revolution in Time,* Cambridge: The Belknap Press of Harvard University Press, 1983.

ing theory of why it is happening, seem to me to be convincing evidence that these curves describe reality and will continue to do so in the future.

A FORECASTING TOOL

If that is true and if the limit for an S-curve can be predicted, then the S-curve can yield valuable insights. If we can define important performance parameters, trace the early days of progress of these parameters versus the effort to make the progress, and develop a point of view on what the limits of these performance parameters are, then we have a basis for foreseeing how much further current products can be improved and how much effort it will take to get them to higher levels of performance (see Appendix 2). If S-curves of potential competitors are also sketched, then we can gain some insight about them. Equally, it will give insights about how products will fare in the future, what new products to try to develop, and how much effort will be required to develop them.

S-curves have been constructed for electric power technology, the accuracy of clocks, increasing the efficiency of electric light bulbs, ammonia production, drug dosage, telecommunications band widths, organic insecticides and software and dozens of other technological developments. I've even seen one for "travel," which includes walking, wagons, railroads, cars and airplanes and covers a long period of time. But in all these cases what we want to know is the relationship between effort put in and results achieved. You might think you should be plotting results against the amount of time involved. But that would be an error. It is not the passage of time that leads to progress, but the application of effort. If we plotted our results versus time, we could not by extrapolation draw any conclusion about the future because we would have buried in our time chart implicit assumptions about the rate of effort applied. If we were to change this rate, it would increase or decrease the time it would take for performance to improve. People frequently make the error of trying to plot technological progress versus time and then find the predictions don't come to pass. Most of the reason for this is not the difficulty of predicting how the technology will evolve, since we have found the S-curve to be rather stable, but rather predicting the rate at which competitors will spend money to develop the technology. The forecasting error is a result of bad competitive analysis, not bad technology analysis.

Thus, it might appear that a technology still has great potential but in fact what is fueling its advance is rapidly increasing amounts of investment. Gordon Moore, President of Intel, says that the density of circuits on a chip will double every two years. That has been true, but how long will it continue to be true? If we plot progress against time, it looks as if progress is getting more rapid. But if we plot chip density as a function of effort, there are signs that the technology is beginning to approach its limit. The rate of effort (dollars per year) has been increasing even more quickly than chip density. This means it is getting more and more expensive to develop each new generation of electronic memories. It took about $100 million to develop the future workhorse of the personal computer—the 256K RAM. What will it take to develop the next generation RAM—the million bit (or megabit) RAM? Probably a good deal more than $100 million.

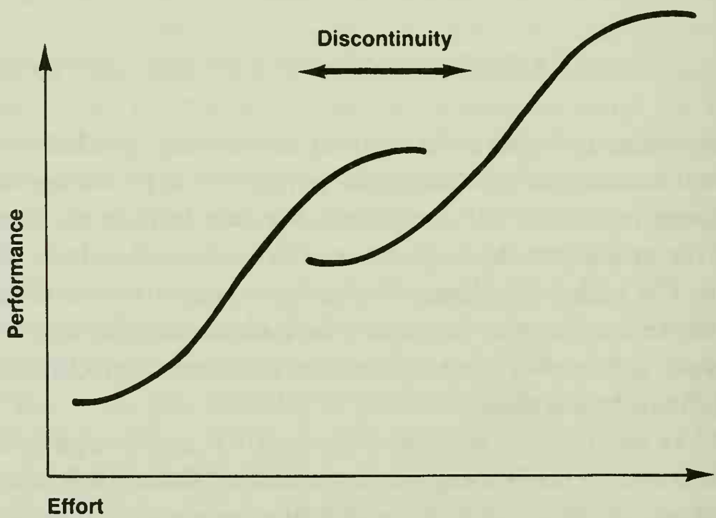
STRATEGIC MANAGEMENT OF DISCONTINUITIES

Ascending an S-curve is something like driving up a hill. You often see caution signs saying 10 percent or 30 percent grade, depending on the hill's steepness. We can think of the slope of the graph just the way we can think about the slope of a hill. The higher the slope, the more productive we are. Thus, a convenient way to pinpoint where we are on our curve of results versus efforts is to talk about the slope or productivity of the technical effort.

At the start of the curve we need to put in significant effort before we can expect to see results. Once the learning is done, we begin to make significant progress for very little expenditure of effort. That usually does not last too long—perhaps a few years. At some point we begin to approach the limits of the technology and we start to run out of steam. Then the question is might there be another way to deliver the desired performance to customers. Some new technology which, though still undeveloped, might eventually outperform the current one, which is increasingly resisting improvement?

But too often we don't ask those questions. Behind conventional management wisdom is the implicit assumption that the more effort put in, the more progress that results. In fact, this is only the case in the first half of the S-curve. In the other half it is wrong. To compound things, it is hard to see what is happening when it is happening because most companies do not keep records of technological productivity.

S-curves almost always come in pairs (Exhibit 12). The gap between the pair of S-curves represents a discontinuity—



12 S-Curves Almost Always Appear in Pairs.
Together they represent a discontinuity—when one technology replaces another.

a point when one technology replaces another. For example, when solid-state electronics replaced vacuum tubes.

In fact, rarely does a single technology meet all customers' needs. There are almost always competing technologies, each with its own S-curve. Thus, in reality, there may be three or four or more technologies involved in a battle, some on defense and some on offense. Often several new technologies vie with each other to replace an old technology in a market segment—for example, the way compact disc players are competing with advanced tape decks and super-refined turntables for a share of the home stereo market. Deciphering a discontinuity's S-curves when all this is happening is very difficult. Not surprisingly, a friend of mine in the pharmaceutical industry says these periods of discontinuity are "chaos." Indeed, they are.

Companies that have learned how to cross technological discontinuities have escaped this trap. They invest in research in order to know where they are on relevant S-curves and know what to expect from the beginning, the middle and the end of these curves. A few draw very precise S-curves, but it's often enough just to know the general dimensions and limits and accept the implications.

This is what I referred to in the opening chapter as the fourth era in the management of technology—the management of discontinuities. Most companies are in the third era, the one I labeled as "strategic" management of technology. They have become very sophisticated at massaging the shape of the curve, making it steeper by developing new products and processes faster than their competitors. For example, my partner Ed Krubasik in McKinsey's Munich office has studied a number of cases involving faster than normal development for individual products. His hypothesis is that the need to cut development times depends on both the costs of development and the profits that might be missed if development were delayed. Often one or both of these can be substantial. He studied several cases to illustrate his point: The IBM PC, the Boeing 767, the Canon PC 10, and Northern Telecom's "digital switch" (solid-state digital telephone exchange). All had either high technology development costs (Boeing), or high market opportunity costs (IBM PC, Canon PC-10), or both (Northern Telecom). What he found was that the cost of being late to the market overwhelmed the cost increases for accelerating development. In each case, special tactics were used to speed up development.

Design and engineering were done as far upstream as possible. For example, Boeing identified, designed and tested things such as composite tail and wing materials before it went to prototype, even to the point of doing wind tunnel tests. In the cockpit Boeing tested several key components as back-up systems in its 737s before going to final design.

Some companies have also learned how to share the development of technology and products. For its personal computer, IBM bought its monitor from Matsushita, its floppy disk from Tandon, its microprocessor from Intel, its printer from Epson, and its operating system from Microsoft. All companies make extensive use of external suppliers, but not too many manage them efficiently. Yet this is a fertile area for saving time and raising quality control. The same is true for customers. The better companies collaborate as much as they can, and often put specialists on site at their customers' plants.

Administrative procedures can also be eliminated to speed up development. Most new IBM products go through a rigorous eight-phase process. But the IBM PC bypassed this, reporting directly to Chairman John Opel. Companies are learning how to use computers and improved communications to speed up development.

Perhaps most important, companies have learned that in order to be fast to the market, they must invest in understanding the science that supports the base of the S-curve. Too many companies develop products empirically. They know things work, but not why they work. They rush through the engineering and then hit some major problem that requires an understanding of the supporting science which they don't have. There is no base for understanding limits, anticipating progress or fixing performance problems as they inevitably occur when a product is developed too quickly. A thousand engineers get held up while frantic calls go out to basic research.

The RAND Corporation, a government-sponsored think tank headquartered in Santa Monica, California, looked into the cost overruns of "Pioneer Process Plants" in 1981.* They noted:

1. Severe underestimation of capital costs is the norm for all advanced technologies.
2. The factors that account for poor cost estimates and poor performance can largely be identified early in the development of the technology long before major expenditures have been made for detailed engineering, much less construction.
3. Seventy-five percent of the cost variance can be attributed to insufficient technical information *before* the project began.

Said more simply, companies that get products to market ahead of their competitors still don't take shortcuts: they do

* Merrow, E.W., et al., "Understanding Cost Growth and Performance Shortfalls in Pioneer Process Plants," RAND, 1981.

their research first, then tackle engineering. As a result their S-curves are steeper than other companies trying to develop the same technology.

But it is important to understand that all of these things, while they can be used to improve the productivity of R&D expenditures and get new products to market fast, can be futile. As futile as NCR trying to sell its new electro-mechanical cash register to customers when they were awaiting electronic ones. None of these efforts will save a company from a new technology.

EFFICIENCY VERSUS EFFECTIVENESS

Superb "third era" management of technology has one major problem: it focuses on efficiency when companies need to be concerned with effectiveness. Effectiveness is set when a company determines *which* S-curve it will pursue (e.g., vacuum tubes or solid state). Efficiency is the slope of the present curve. Effectiveness deals with sustaining a strategy–efficiency with the present utilization of resources. Moving into a new technology almost always appears to be less efficient than staying with the present technology because of the need to bring the new technology up to speed. The cost of progress of an established technology is compared with that of one in its infancy, even though it may eventually cost much less to bring the new technology up to the state of the art than it did to bring the present one there. To paraphrase a comment I've heard many times at budget meetings: "In any case the new technology development cost is above and beyond what we're already paying. Since it doesn't get us any further than we presently are, it cannot make sense." The problem with that argument is that someday it will be ten or twenty or thirty times more efficient to invest in the new technology, and it will outperform the existing technology by a wide margin.

There are many decisions that put effectiveness and efficiency at odds with each other, particularly those involving resource allocation. This is one of the toughest areas to come to grips with because it means withdrawing resources from the maturing business. What makes this decision so difficult is that it is being made from the inside and resources will have to be withdrawn from businesses that in many cases sired the CEO, or from the division that has been the major building block for the company and has the strongest political ties to its current management.

In addition, many companies have management policies that, interpreted literally, impede moving from one S-curve to another. For example, "Our first priority will be to protect our existing businesses." Or "We will operate each business on a self-sustaining basis; each will have to provide its own cash as well as make a contribution to corporate overhead." These rules are established either in a period of relaxed competition or out of political necessity.

The fundamental dilemma is that it always appears to be more economic to protect the old business than to feed the new one at least until competitors pursuing the new approach get the upper hand. Conventional financial theory has no practical way to take account of the opportunity cost of not investing in the new technology. If it did, the decision to invest in the present technology would often be reversed.

In military strategy the conventional wisdom for ground-based warfare is that the defender has the advantage. He is up on the hill with a clear view and can see the enemy coming. Military strategists from Von Clausewitz to Liddell Hart, as well as practitioners like Eisenhower and Rommel, all felt that if you wanted to take a hill held by a defender you needed to show up with three times his number or risk getting slaughtered. The spirit of that point of view is carried forward into business as well, with the belief that the defender, the competitor with the largest market share, the most knowledge of production processes and distribution, will have the advantage in combat in the marketplace. I believe the reverse is true. The defender is at an inherent disadvantage. He may not even know he is being attacked until the attack is well along. The attacker can hide in a niche. He is often more powerful than he appears, and more motivated.

A few years ago we polled the R&D vice presidents of 250 of America's largest firms who constitute the Industrial Research Institute. One of the most important findings was their belief that on average U.S. companies could double their R&D productivity. Such is the untapped potential of defenders that half the gain, they believed, could come from a more effective choice of projects and the other half from improvements in the work performed.

Our work at McKinsey corroborates this. When we analyze the technical spending of large companies, it is not uncommon to find 80 percent of the effort going to the defense of products that are more important for what they have contributed in the past than for what they are going to contribute to the future. This consumes funds that could be spent on the technical or market exploration of fields with higher potential or higher productivity. If there is usually a 5-to-1 difference in productivity between investments in emerging and mature technologies, then shifting just one dollar in five currently invested in the mature technology into the newer one would almost double the results.

In the 1K random-access memory, which was the first built, the productivity differences between the emergent and adolescent stages were on the order of 19 to 1. At McKinsey we have seen differences in productivity in electronics technology on the order of 20 to 1, even 30 to 1. That is 3,000 percent difference because of the basic choice of technology. People wondered how in the late 1950s tiny Texas Instruments could compete with giants like Westinghouse or Sylvania. Try multiplying TI's size back then by 30. By being on the right technological S-curve, TI compensated for much of its size and market power disadvantage.

In pharmaceuticals today, barely 10 percent of the R&D funding goes toward the newer biological approach to creating drugs. Yet my guess is that at least half the total innovations will come out of that area, so much higher is its potential.

Managers often talk about improving the productivity of a plant or of sales, but it is not often they expect anything higher than a 10 percent or 15 percent gain. We are talking 100 percent and 500 percent differences in technical productivity between competitors because one made the right technological choice and the other did not. No other area of management can match technology when it comes to potential gains in worker output.

Even if a defender succeeds in managing his own S-curve better, chances are he will not be able to raise his efficiency by more than, say, 50 percent. Not much use against an attacker whose productivity might be climbing ten times faster because he has chosen a different S-curve. All too frequently the defender believes his productivity is actually higher than his attacker's and ignores what the attacker potentially may have to offer the customer. Defenders and attackers often have a different perspective when it comes to judging productivity. For the attacker, productivity is the improvement in performance of his new product over his old product divided by the effort he puts into developing the new product. If his technology is beginning to approach the steep part of its S-curve, this could be a big number. The defender, however, observes the productivity through the eyes of the market, which may still be treating the new product as not much more than a curiosity. So in his eyes the attacker's productivity is quite low. We've seen this happen time and again in the electronics industry. Products such as microwaves, audio cassettes and floppy discs failed at first to meet customer standards, but then, almost overnight, they set new high-quality standards and stormed the market.

Even if the defender admits that the attacker's product may have an edge, he is likely to say it is too small to matter. Since the first version of a wholly new product is frequently just marginally better than the existing product, the defender often thinks the attacker's productivity is lower, not higher than his own. The danger comes in using this erroneous perception to figure out what is going to happen next. Too often defenders err by thinking that the attacker's second generation new product will require enormous resources and result in little progress. We know differently. We know from the mathematics of adolescent S-curves that once the first crack appears in the market dam, the flood cannot be far behind. And further, it won't cost nearly as much since the first product has absorbed much of the start-up costs. No doubt this will be a big shock to the defender who will tell the stock market analysts, "Well, the attacker was just lucky. There was nothing in his record to suggest he could have pulled this thing off." All true. From the defender's viewpoint there was nothing in the attacker's record to suggest that a change was coming. But the underlying forces were at work nevertheless, and in the end they appeared.

Reallocating resources is thus a painful business. At times any decision that top management makes, any action they take may well be viewed as contrary to the company's best interests. Often CEOs will be roundly criticized by outsiders for venturing into new areas where they lack skills and for forsaking the tried and true. But in order to manage a technological discontinuity, that is exactly what they must do—forsake the past by abandoning a technology that, more often than not, has just entered the most productive phase of its S-curve. This dilemma captures the dimensions of the fourth era of managing technology which companies must now enter. It involves knowledge building, analysis, and the calculation of limits. And it involves, indeed requires, the conviction and courage to realize that sometimes it is necessary to cut off your arm.