Microprocessors in 2020

Every 18 months microprocessors double in speed. Within 25 years, one computer will be as powerful as all those in Silicon Valley today

by David A. Patterson

When I first read the table of contents of this special issue, I was struck by how many articles addressed computers in the 21st century in some way. Unlike many other technologies that fed our imaginations and then faded away, the computer has transformed our society. There can be little doubt that it will continue to do so for many decades to come. The engine driving this ongoing revolution is the microprocessor. These silicon chips have led to countless inventions, such as portable computers and fax machines, and have added intelligence to modern automobiles and wristwatches. Astonishingly, their performance has improved 25,000 times over since their invention only 25 years ago.

I have been asked to describe the microprocessor of 2020. Such predictions in my opinion tend to overstate the worth of radical, new computing technologies. Hence, I boldly predict that changes will be evolutionary in nature, and not revolutionary. Even so, if the microprocessor continues to improve at its current rate, I cannot help but suggest that 25 years from now these chips will empower revolutionary software to compute wonderful things.

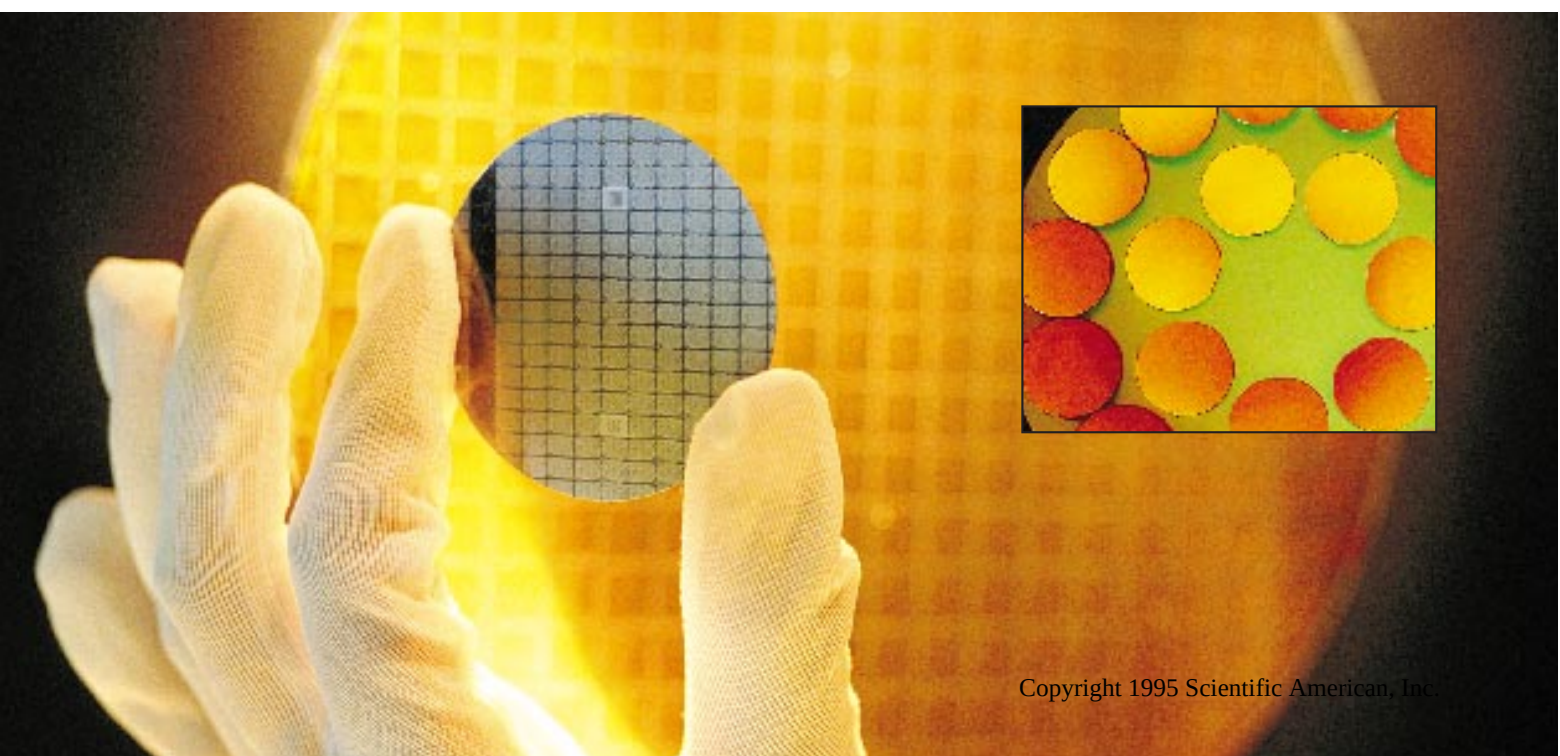
Smaller, Faster, Cheaper

Two inventions sparked the computer revolution. The first was the so-called stored program concept. Every computer system since the late 1940s has adhered to this model, which prescribes a processor for crunching numbers and a memory for storing both data and programs. The advantage in such a system is that, because stored programs can be easily interchanged, the same hardware can perform a variety of tasks. Had computers not been given this flexibility, it is probable that they would not have met with such widespread use. Also, during the late 1940s, researchers invented the transistor. These silicon switches were much smaller than the vacuum tubes used in early circuitry. As such, they enabled workers to create smaller—and faster—electronics.

More than a decade passed before the stored program design and transistors were brought together in the same machine, and it was not until 1971 that the most significant pairing—the Intel 4004—came about. This processor was the first to be built on a single silicon chip, which was no larger than a child's fingernail. Because of its tiny size, it was dubbed a microprocessor. And because it was a single chip, the Intel 4004 was the first processor that could be made inexpensively in bulk.

The method manufacturers have used to mass-produce microprocessors since then is much like baking a pizza: the dough, in this case silicon, starts thin and round. Chemical toppings are added, and the assembly goes into an oven. Heat transforms the toppings into transistors, conductors and insulators. Not surprisingly, the process—which is repeated perhaps 20 times—is considerably more demanding than baking a pizza. One dust particle can damage

CHARLES O'REAR



the tiny transistors. So, too, vibrations from a passing truck can throw the ingredients out of alignment, ruining the end product. But provided that does not happen, the resulting wafer is divided into individual pieces, called chips, and served to customers.

Although this basic recipe is still followed, the production line has made ever cheaper, faster chips over time by churning out larger wafers and smaller transistors. This trend reveals an important principle of microprocessor economics: the more chips made per wafer, the less expensive they are. Larger chips are faster than smaller ones because they can hold more transistors. The recent Intel P6, for example, contains 5.5 million transistors and is much larger than the Intel 4004, which had a mere 2,300 transistors. But larger chips are also more likely to contain flaws. Balancing cost and performance, then, is a significant part of the art of chip design.

Most recently, microprocessors have become more powerful, thanks to a change in the design approach. Following the lead of researchers at universities and laboratories across the U.S., commercial chip designers now take a quantitative approach to computer architecture. Careful experiments precede hardware development, and engineers use sensible metrics to judge their success. Computer companies acted in concert to adopt this design strategy during the 1980s, and as a result, the rate of improvement in microprocessor technology has risen from 35 percent a year only a decade ago to its current high of approximately 55 percent a year, or almost 4 percent each month. Processors are now three times faster than had been predicted in the early 1980s; it is as if our wish was granted, and we now have machines from the year 2000.

Pipelined, Superscalar and Parallel

In addition to progress made on the production line and in silicon technology, microprocessors have benefited from recent gains on the drawing board. These breakthroughs will undoubtedly lead to further advancements in the near future. One key technique is called pipelining. Anyone who has done laundry has intuitively used this tactic. The nonpipelined approach is as follows: place a load of dirty clothes in the washer. When the washer is done, place the wet load into the dryer. When the dryer is finished, fold the clothes. After the clothes are put away, start all over again. If it takes an hour to do one load this way, 20 loads take 20 hours.

The pipelined approach is much quicker. As soon as the first load is in the dryer, the second dirty load goes into the washer, and so on. All the stages operate concurrently. The pipelining paradox is that it takes the same amount of time to clean a single dirty sock by either method. Yet pipelining is faster in that more loads are finished per hour. In fact, assuming that each stage takes the same amount of time, the time saved by pipelining is proportional to the number of stages involved. In our example, pipelined laundry has four stages, so it would be nearly four times faster than nonpipelined laundry. Twenty loads would take roughly five hours.

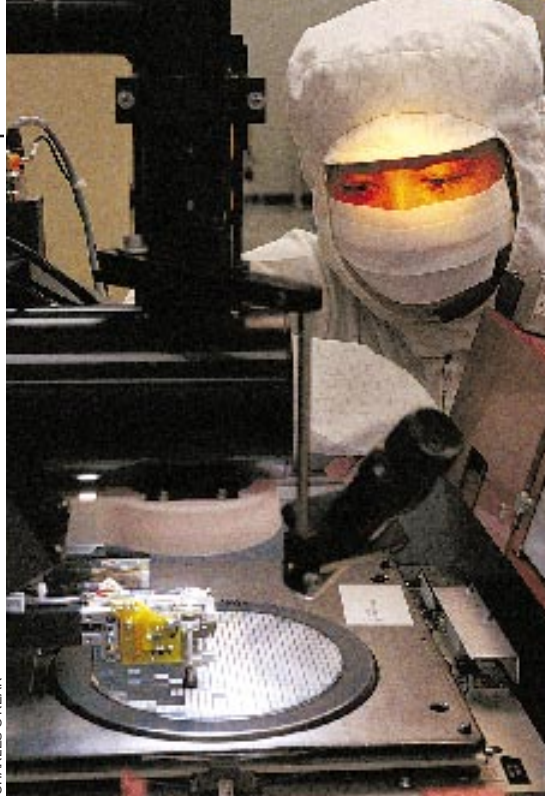
Similarly, pipelining makes for much faster microprocessors. Chip designers pipeline the instructions, or low-level commands, given to the hardware. The first pipelined microprocessors used a five-stage pipeline. (The number of stages completed each second is given by the so-called clock rate. A personal computer with a 100-megahertz clock then executes 100 million stages per second.) Because the speedup from pipelining equals the number of stages, recent microprocessors have adopted eight or more stage pipelines. One 1995 microprocessor uses this deeper pipeline to achieve a 300-megahertz clock rate. As machines head toward the next century, we can expect pipelines having even more stages and higher clock rates.

Also in the interest of making faster chips, designers have begun to include more hardware to process more tasks at each stage of a pipeline. The buzzword "superscalar" is commonly used to describe this approach. A superscalar laundromat, for example, would use a professional machine that could, say, wash three loads at once. Modern superscalar microprocessors try to perform anywhere from three to six instructions in each stage. Hence, a 250-megahertz, four-way superscalar microprocessor can execute a billion instructions per second. A 21st-century microprocessor may well launch up to dozens of instructions in each stage.

Despite such potential, improvements in processing chips are ineffectual unless they are matched by similar gains in memory chips. Since random-access memory (RAM) on a chip became widely available in the mid-1970s, its capacity has grown fourfold every three years. But memory speed has not increased at anywhere near this rate. The gap between the top speed of processors and the top speed of memories is widening.

One popular aid is to place a small memory, called a cache, right on the microprocessor itself. The cache holds those segments of a program that are most frequently used and thereby allows the processor to avoid calling on external memory chips much of the time. Some newer chips actually dedicate as many transistors to the cache as they do to the processor itself. Future microprocessors will allot even more resources to the cache to better bridge the speed gap.

The Holy Grail of computer design is an approach called parallel processing, which delivers all the benefits of a single fast processor by engaging many inexpensive ones at the same time. In our analogy, we would go to a laundromat and use 20 washers and 20 dryers to do 20 loads simultaneously. Clearly, parallel processing is an expensive solution for a small workload. And writing a program that can use 20 processors at once is much harder than distributing laundry to 20 washers. Indeed,



CHARLES O'REAR

CLEAN ROOMS, where wafers are made, are designed to keep human handling and airborne particles to a minimum. A single speck of dust can damage a tiny transistor.

SILICON WAFERS today (*background*) are much larger but hold only about half as many individual chips as did those of the original microprocessor, the Intel 4004 (*foreground*). The dies can be bigger in part because the manufacturing process (*one stage shown in inset*) is cleaner.

The Limits of Lithography

Although I predict that microprocessors will continue to improve rapidly, such a steady advance is far from certain. It is unclear how manufacturers will make tinier, faster transistors in the years to come. The photolithographic methods they now use are reaching serious limits. If the problem is not resolved, the progress we have enjoyed for decades will screech to a halt.

In photolithography, light is used to transfer circuit patterns from a quartz template, or mask, onto the surface of a silicon chip. The technique now fashions chip features that are some 0.35 micron wide. Making features half as wide would yield transistors four times smaller, since the device is essentially two-dimensional. But it seems impossible to make such tiny parts using light; the light waves are just too wide. Many companies have invested in finding ways to substitute smaller x-rays for light waves. To date, however, x-rays have not succeeded as a way to mass-produce state-of-the-art chips.

CHARLES O'REAR

PHOTOMASKS are reduced and projected onto silicon wafers to make circuits.

Other proposals abound. One hope is to deploy the electron beams used to create quartz masks to pattern silicon wafers. The thin stream of charged particles could trace each line in a circuit diagram, one by one, directly onto a chip. The catch is that although this solution is feasible, it is unreasonably slow for commercial use and would therefore prove costly. Compared with photolithography, drawing with an electron beam is analogous to rewriting a letter by hand instead of photocopying it.

Technical hurdles aside, any improvements in microprocessors are further threatened by the rising cost of semiconductor manufacturing plants. At $1 billion to $2 billion, these complexes now cost 1,000 times more than they did 30 years ago. Buyers and sellers of semiconductor equipment follow the rule that halving the minimum feature size doubles the price. Clearly, even if innovative methods are found, the income generated by the sale of smaller chips must double to secure continued investments in new lines. This pattern will happen only by making more chips or by charging more for them.

Today there are as many companies that have semiconductor lines as there are car companies. But increasingly few of them can afford the multibillion-dollar cost of replacing the equipment. If semiconductor equipment manufacturers do not offer machinery that trades off, say, the speed of making a wafer for the cost of the equipment, the number of companies making state-of-the-art chips may shrink to a mere handful. Without the spur of competition, once again, the rapid pace of improvement may well slow down. —*D.A.P.*

the program must specify which instructions can be launched by which processor at what time.

Superscalar processing bears similarities to parallel processing, and it is more popular because the hardware automatically finds instructions that launch at the same time. But its potential processing power is not as large. If it were not so difficult to write the necessary programs, parallel processors could be made as powerful as one could afford. For the past 25 years, computer scientists have predicted that the programming problems will be overcome. In fact, parallel processing is practical for only a few classes of programs today.

In reviewing old articles, I have seen fantastic predictions of what computers would be like in 1995. Many stated that optics would replace electronics; computers would be built entirely from biological materials; the stored program concept would be discarded. These descriptions demonstrate that it is impossible to foresee what inventions will prove commercially viable and go on to revolutionize the computer industry. In my career, only three new technologies have prevailed: microprocessors, random-access memory and optical fibers. And their impact has yet to wane, decades after their debut.

Surely one or two more inventions will revise computing in the next 25 years. My guess, though, is that the stored program concept is too elegant to be easily replaced. I believe future computers will be much like machines of the past, even if they are made of very different stuff. I do not think the microprocessor of 2020 will be startling to people from our time, although the fastest chips may be much larger than the very first wafer, and the cheapest chips may be much smaller than the original Intel 4004.

IRAMs and Picoprocessors

Pipelining, superscalar organization and caches will continue to play major roles in the advancement of microprocessor technology, and if hopes are realized, parallel processing will join them. What will be startling is that microprocessors will probably exist in everything from light switches to pieces of paper. And the range of applications these extraordinary devices will support, from voice recognition to virtual reality, will very likely be astounding.

Today microprocessors and memories are made on distinct manufacturing lines, but it need not be so. Perhaps in the near future, processors and memory will be merged onto a single chip, just as the microprocessor first merged the separate components of a processor onto a single chip. To narrow the processor-memory performance gap, to take advantage of parallel processing, to amortize the costs of the line and simply to make full use of the phenomenal number of transistors that can be placed on a single chip, I predict that the high-end microprocessor of 2020 will be an entire computer.

Let's call it an IRAM, standing for intelligent random-access memory, since most of the transistors on this merged chip will be devoted to memory. Whereas current microprocessors rely on hundreds of wires to connect to external memory chips, IRAMs will need no more than computer network connections and a power plug. All input-output devices will be linked to them via networks. If they need more memory, they will get more processing power as well, and vice versa—an arrangement that will keep the memory capacity and processor speed in balance. IRAMs are also the ideal building block for parallel processing. And because they would require so few external connections, these chips

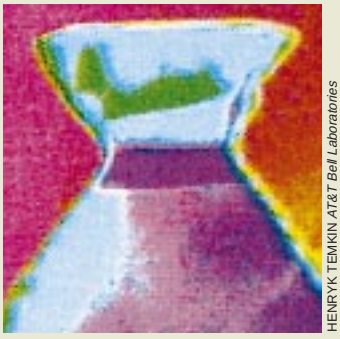
And After 2020?

With decades of innovative potential ahead of them, conventional microelectronic designs will dominate much of the 21st century. That trend does not discourage many laboratories from exploring a variety of novel technologies that might be useful in designing new generations of computers and microelectronic devices. In some cases, these approaches would allow chip designs to reach a level of miniaturization unattainable through anything like conventional lithography techniques. Among the ideas being investigated are:

▪ **Quantum dots and other single-electron devices.** Quantum dots are molecular arrays that allow researchers to trap individual electrons and monitor their movements. These devices can in theory be used as binary registers in which the presence or absence of a single electron is used to represent the 0 or 1 of a data bit. In a variation on this scheme, laser light shining on atoms could switch them between their electronic ground state and an excited state, in effect flipping the bit value.

One complication of making the transistors and wires extremely small is that quantum-mechanical effects begin to disrupt their function. The logic components hold their 0 or 1 values less reliably because the locations of single electrons become hard to specify. Yet this property could be exploited: Seth Lloyd of the Massachusetts Institute of Technology and other researchers are studying the possibility of developing quantum computing techniques, which would capitalize on the nonclassical behavior of the devices.

▪ **Molecular computing.** Instead of making components out of silicon, some investigators are trying to develop data storage systems using biological molecules. Robert L. Birge of Syracuse University, for example, is examining the computational potential of molecules related to bacteriorhodopsin, a pigment that alters its configuration in response to light. One advantage of such a molecule is that it could be used in an optical computer, in which streams of photons would take the place of electrons. Another is that many of these molecules might be synthesized by microorganisms, rather than fabricated in a factory. According to some estimates, photonically activated biomolecules could be linked into a three-dimensional memory system that would have a capacity 300 times greater than today's CD-ROMs.



HENRYK TEMKIN *AT&T Bell Laboratories*

QUANTUM DOT **(purple)** ***in this semiconductor structure traps electrons.***

▪ **Nanomechanical logic gates.** In these systems, tiny beams or filaments only one atom wide might be physically moved, like Tinkertoys, to carry out logical operations [see "Self-Assembling Materials," by George M. Whitesides, page 146].

▪ **Reversible logic gates.** As the component density on chips rises, dissipating the heat generated by computations becomes more difficult. Researchers at Xerox PARC, the IBM Thomas J. Watson Research Center and elsewhere are therefore checking into the possibility of returning capacitors to their original state at the end of a calculation. Because reversible logic gates would in effect recapture some of the energy expended, they would generate less waste heat.

—The Editors

could be extraordinarily small. We may well see cheap "picoprocessors" that are smaller than the ancient Intel 4004. If parallel processing succeeds, this sea of transistors could also be used by multiple processors on a single chip, giving us a micromultiprocessor.

Today's microprocessors are almost 100,000 times faster than their Neanderthal ancestors of the 1950s, and when inflation is considered, they cost 1,000 times less. These extraordinary facts explain why computing plays such a large role in our world now. Looking ahead, microprocessor performance will easily keep doubling every 18 months through the turn of the century. After that, it is hard to bet against a curve that has outstripped all expectations. But it is plausible that we will see improvements in the next 25 years at least as large as those seen in the past 50. This estimate means that one desktop computer in 2020 will be as powerful as all the computers in Silicon Valley today. Polishing my crystal ball to look yet another 25 years ahead, I see another quantum jump in computing power. The implications of such a breathtaking advance are limited only by our imaginations. Fortunately, the editors have asked others to ponder the possibilities, and I happily pass the baton to them.

The Author

DAVID A. PATTERSON has taught since 1977 at the University of California, Berkeley, where he now holds the E. H. and M. E. Pardee Chair in Computer Science. He is a member of the National Academy of Engineering and is a fellow of both the Institute of Electrical and Electronic Engineers and the Association for Computing Machinery. He has won several teaching awards, co-authored five books and consulted for many companies, including Digital, Intel and Sun Microsystems. His current research is on large-scale computing using networks of workstations.

Further Reading

MICROPROCESSORS: FROM DESKTOPS TO SUPERCOMPUTERS. F. Baskett and J. L. Hennessy. *Science,* Vol. 261, pages 864–871; August 13, 1993.

COMPUTER ORGANIZATION AND DESIGN: THE HARDWARE/SOFTWARE INTERFACE. J. L. Hennessy and D. A. Patterson. Morgan Kaufmann Publishers, 1994.

COMPUTER ARCHITECTURE: A QUANTITATIVE APPROACH. Second edition. D. A. Patterson and J. L. Hennessy. Morgan Kaufmann Publishers, 1995.

COMPUTING PERSPECTIVES. M. V. Wilkes. Morgan Kaufmann Publishers, 1995.

Follow the reference on the World Wide Web http://cra.org:80/research.impact/ and look under "RISC" to learn more about the rapid rise in processor performance.