Chapter 3

# High-Performance Computers: Technology and Challenges

## Computers and the R&D Process

Scientists use the theories and techniques of mathematics for building and describing models in logical ways and for calculating the results they yield. As early as the third century B. C., the Alexandria scholar Eratosthenes estimated the circumference of the earth to an accuracy within 5 percent of what we now consider to be the correct figure. He did so by making assumptions about the nature of the physical universe, making measurements, and calculating the results. [1] In essence, he did what modern scientists do. He constructed a hypothetical model that allowed him to apply mathematical tools—in this case, trigonometry and arithmetic—to data he collected.

Scientific models are used both to test new ideas about the physical universe and to explore resu1ts and conclusions based on those models. Eratosthenes discovered a new ''fact' '—the size of the earth. Had his calculations, instead, confirmed a result already discovered by some other means, he would have accomplished a different research purpose; he would have provided evidence that the model of the universe was correct. Had they differed with known fact, he would have had evidence that the model was incorrect. Science advances, step by step, through a process of building models, calculating results, comparing those results with what can be observed and, when observations differ, revising the models.

### Modes Of Research Computing

Just as mathematics is central to science, computers have become basic instruments of research to modern science and play a wide variety of roles. Each of the roles is based on mathematical modeling, using the interactive solution of thousands of equations.

#### To Perform Complex Calculations

Sometimes the basic mathematics and structure of a physical process are well known—the equations that describe the flow of air around a solid object, for example. Researchers may wish to calculate the results of this process in experimental designs such as a new aircraft wing or the shape of an automobile. Calculating results from flow equations are enormously time-consuming even on the most powerful computers of today. Scientists must simplify these problems to fit the capabilities of the computers that are available. They sacrifice accuracy and detail in their model to achieve computability.

#### To Build New Theories and Models

At other times, researchers seek to understand the dynamics of a process, like the aging of a star or formation of a galaxy. They create computer models based on theories and observe how the behavior of those models do or do not correspond to their observations.

#### To Control Experimental Instruments and Analyze Data

Most modern scientific instruments have some computational power built in to control their performance and to process the measurements they make. For many of these, from the largest particle accelerators or space platforms to more modest instruments, the computer has become an integral and indispensable part.

Such research instruments generate enormous flows of information-some at rates up to several trillion units (terabits) a day. Unpackaging the data flow, identifying the elements, and organizing those data for use by scientists is, itself, a sizable computational task. After the initial steps, still more computer power is needed to search this mountain of data for significant patterns and analyze their meanings.

#### To Better Understand and Interact With Computer Results

At the most basic level, computers produce numbers; but numbers usually represent a physical object or phenomenon-the position of an atom in a protein molecule, the moisture content in a cloud, the stress in an automobile frame, or the behavior of an explosive. To make sense to researchers, the streams of numbers from a computer must be

[1] Thomas S Kuhn, *The Copernican Revolution* (Cambridge, MA: Harvard University Press, 1985), p. 274.

converted to visual displays that are easier to understand when seen by the eye. Researchers are now concentrating on visualization-pictorial displays that incorporate images, motion, color, and surface texture to depict characteristics of an analysis on a computer screen.

Some researchers are exploring more advanced techniques that use other senses such as sound and touch to convey results to the human mind. By incorporating all of these technologies, they may eventually be able to create what is called ‘‘virtual reality, in which a scientist equipped with the proper gear could interact directly with a model as though he or she were standing in the midst of the phenomenon that was modeled. A biochemist could ‘‘walk’ around and about a protein molecule, for example, and move atoms here and there, or a geologist could explore the inside of an active volcano.

### To Provide “Intelligent” Assistance

Computer operations are not restricted to only computational operations on numbers. The popularity of word processors shows that computers can manipulate and perform logical operations on symbols, whether they represent numbers or not. Experts in the ‘‘artificial intelligence’ community have been exploring how computers can assist researchers in ways other than direct computation of results. They have worked on systems that can prove mathematical theorems or perform tedious manipulations of algebraic expressions, systems that help chemists find new forms of molecules, and natural language inquiry systems for databases.

A national research and educational network (NREN) would create a critical need for such help in the future so that scientists are not overwhelmed by the complexity and amount of information available to them. New tools such as “knowbots’’—small autonomous programs that would search databases throughout the network for information needed by the researcher-have been proposed.

### *Implications for Federal Programs*

The traditional view of the ‘ ‘scientific computer’ as one specifically intended for high-speed arithmetic computation is changing as researchers use computers for an increasingly rich variety of tasks. Any Federal initiative supporting computational science must create an environment that supports a wide variety of machines with improved capabilities, many of which serve specialized user communities.

Numerical computation is still critically important, but so are applications such as database manipulation, artificial intelligence, image production, and on-line control of experimental instruments. Even the design of computers meant to do numerical calculations is becoming more specialized to address specific types of problems.

The NREN is a crucial element of efforts to make high-performance computing widely available to the U.S. research community. Members of research groups who need these specialized computers are widely scattered throughout the country, and so are the computers they need.

## The Evolution of Computer Technology

### *Government and Computer R&D*

Like much of the new electronics technology of the day, computers in large measure grew out of work done during World War II for defense research programs. After the war, many engineers and scientists who staffed those programs took their knowledge into the private sector to begin the commercial U.S. computer industry.

The Federal Government remains a major purchaser, user, and force in shaping computer technology. Its influence is particularly strong in scientific computing; many computational researchers either work for the government in national laboratories or are substantially funded by government agencies. The computing needs of the defense agencies, and the weapons programs of the Department of Energy (earlier the Atomic Energy Commission (AEC)), demanded continual advancement of the speed and power of scientific computing.

Computers that meet the specifications of scientific users were not, until recently, commercially successful or widely available. As a result, Federal agencies needing these large scientific machines had to fired their development. Control Data’s 6600 computer in the mid- 1960s was among the first large scientific machines designed for national defense needs to be marketed successfully in the private sector.

Even though scientific computers were not originally successful in the nongovernment market, their technology was. The ''Stretch' computer, designed and built by IBM for the AEC, provided many innovations that were later used in the design of the IBM 360 series that was the basic IBM product line for over a decade. Federal science agencies such as the National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), and the Office of Naval Research (ONR) have also contributed over the years to the development of computer architecture through their computer science and engineering research programs.

The government role in support of basic and applied research in computing and in testing prototype machines and making them available to researchers is critical to the well-being of small specialized firms in high-performance computing.

Government support for research in computer architecture has gone through cycles. In the early days, it was in research laboratories that computer scientists first developed many of the architectural concepts that formed the basis for general purpose computers. As computers became more complex and their manufacture a more refined art, academic research on computer design waned. Perhaps the decreased interest in architecture research resulted from the notion at that time that the major computer design issues had been settled and the development of new generations of machines should be left to the industry. The academic research that continued was mostly paper-and-pencil design simulated on conventional computers.

During the last decade, advances in microelectronics created opportunities to explore radical new designs with relatively inexpensive off-the-shelf chips from manufacturers, or custom designs. Experts were predicting the end of performance improvements that could be wrung from traditional design concepts, while the costs for coaxing performance improvements were increasing dramatically. As a result, computer scientists and engineers are again exploring alternate approaches, and academic research has now returned to the development and testing of prototypes, this time in cooperation with industry. Now, as then, the basic question is whether these experimental designs are more efficient and effective for performing specific types of calculations.

Computer scientists and engineers basically look in three directions to improve the efficiency and increase the speed of computers:

1. the fundamental technology of the computer components;
2. the architecture of the computer; and
3. the software programs and algorithms to instruct and control the computers.

These three areas of investigation are distinct fields of research, but they have an important influence on each other. New devices allow computer designers to consider different approaches to building computers, which, in turn, can lead to new ways of programming them. Influences can just as easily go the other way: new software techniques can suggest new machine architectures. One of the problems with introducing radically new types of computers into common use is that entirely new theories of programming must be developed for them, whereas software techniques for traditional machines have taken place over 40 or 50 years of development and refinement.

### Fundamental Technologies

Basically, computers are complex assemblies of large numbers of essentially similar building blocks. These building blocks—all of which are generally different types of logical switches that can be set in one of two states (on-off)--are combined to form the memory, registers, arithmetic units, and control elements of modern digital computers (see box C). The advance of computer technology at this level can be seen as the clustering of more and more of these basic switches into increasingly smaller, faster, cheaper, and more reliable packages.

***Integrated*** Circuits—Electrical engineers predict that, by 2000, chip manufacturers will be able to put over one billion logic gates (switches) on a single chip. Some silicon chips already contain more than a million gates. This level of complexity begins to allow producers to put huge computational power on one processor chip. By the end of the decade, it is expected that a single chip will have the complexity and the power of a modern supercomputer, along with a significant amount of memory.

This trend is influencing research in computer design. Computer scientists and engineers use the term ''architecture' to describe the art of arranging the flows of data and the detailed logical processes within the computers they design. Given the com-

## ***Box** C—The Building Blocks of Modern Computer Hardware*

From electro-mechanical relays to vacuum tubes to silicon-based very-large-scale integrated circuits, the electronic technologies that form the basic components of computers have steadily and rapidly advanced year by year since the 1940s. One measure of improvement is the number of transistors (the basic building block of logic and memory) that can be placed on a chip. Increase in transistor density is expected to continue throughout the coming decade, although “traditional” silicon technology, the basis of microelectronics for the last few decades may begin reaching its maximum cost/performance benefit, It may become too costly to derive future performance advancements out of silicon.

In the past, as each type of technology—mechanical switches, vacuum tubes, and transistors-reached its limits, a new technology has come along that allowed information technology to continue improving; this phenomenon is likely to continue. Researchers are exploring several basic technologies that, if successful, could continue these rates of growth, not only through this decade, but well into the next century.[1]

### Gallium Arsenide Compounds

Gallium Arsenide (GaAs) is a compound with semiconductor properties similar to, but in some ways superior to, silicon. Spurred in part by interest from the Department of Defense, researchers have developed GaAs to the point where such devices are being produced for commercial application. But will it ever be cost-effective to manufacture devices complex enough and in quantities sufficient to build full-scale computers in a cost-effective way? Some manufacturers are trying.

Cray Computer Corp. (CCC), a separate company spun off from its parent Cray Research, and Convex Computers-a manufacturer of entry-level supercomputers-are attempting to use GaAs-based components for their new machines. Although offering much greater speeds for the machine, these components have proved to be difficult to manufacture and to assemble into a large-scale mainframe. Their efforts are being watched closely. Some experts think that some of these manufacturing difficulties are inherent and that GaAs will remain a valuable but expensive ‘‘niche’ technology, possibly useful for high-speed and costly applications, but not serving as the ‘‘workhorse’ all-purpose replacement for silicon in everyday applications.[2]

### Superconductivity

For years it has been known that some materials attain a state known as “superconductivity” when cooled sufficiently. A superconductive material essentially transmits electricity without (or with low) resistance. Using superconductivity, a switch known as a “Josephson Junction” (JJ) can be built that could, in theory, serve as the basis of computer logic and memory.

The problem has been that ‘‘sufficiently cooled” has meant very cold indeed, nearly the temperature of liquid helium, only 4 degrees Kelvin.[3] Although it is possible to attain these temperatures, it requires extensive and complex apparatus either for refrigerating or for using liquid helium, a very temperamental substance to deal with. Problems with reliably manufacturing JJs have also been difficult to solve. Because JJs could move computer capabilities beyond silicon limits if these problems were solved, some manufacturers, particularly the Japanese, have continued to explore low-temperature superconductivity.

Within the last few years, however, the discovery of materials that exhibit superconductivity at higher temperatures has led to a renewed interest in the JJ.[4] ‘‘High temperature ‘‘ is still very cold by normal standards, around 50 to 100 degrees Kelvin, but it is a temperature that is much more economical to maintain. Significant materials problems still confound attempts to manufacture JJs reliably and in the bulk necessary to manufacture computers. However, investigators have just begun exploring this technology, and many of them expect that these

[1] U.S. Congress, Office of Technology Assessment, Microelectronics *Research and Development—Background Paper,* OTA-BP-CIT-40 (Washington, DC: U.S. Government Printing Office, March 1986).

[2] Marc H. Brodsky, ‘‘Progress in Gallium Arsenide Semiconductors, ’ *Scientific American,* February 1990, pp. 68-75.

[3] Kelvin is a unit of measurement that uses as its reference, “absolute Zero, ” the coldest temperature that matter can theoretically attain. In comparison, zero degrees Centigrade, the temperature at which water freezes, is a warm 273 degrees Kelvin.

[4] U.S. Congress, Office of Technology Assessment, *Commercializing High-Temperature Superconductivity,* OTA-ITE-388 Washington, DC: U.S. Government Printing Office, August 1988).

problems will be solved, in part because of the potential importance of the technology if it can be tamed. It has been suggested that Japanese manufacturers continue to work on low-temperature prototypes in order to gain experience in designing and building JJ-based computers that could be useful if and when high-temperature technology becomes available.

### other Advanced Technologies

Researchers are also investigating other promising technologies, such as ‘‘optical switching’ devices. Fiber optics already offers significant advantages as a communication medium, but signals must be converted back to electrical form before they can be manipulated. It might be attractive in terms of speed and economy if one could handle them directly in the form of light.

Other researchers are working on so-called “quantum effect” devices. These devices use silicon—and in some cases (GaAs-materials, but take advantage of the quantum, or wave-like, behavior of electrons when they are confined in very small areas (say, on the order of 100 atoms in diameter.)[5] Again, problems of manufacturing, particularly devices as small as this, present major difficulties to be overcome.

[5] Henry I. Smith and Dimitra A. Antoniadis, ‘‘Seeking a Radically New Electronics,’ *Technology Review,* April 1990, pp. 27-39.

plexity that modern chips can embody, a chip designer can use them to build bigger, more elaborate constructs. Such a designer might be thought of more as a ‘‘city planner’—someone who arranges the relationships between much larger structures and plans the traffic flow among them.

Computer design is helped considerably by modern technology. First, through use of automated design and ‘‘chip foundries for producing customized chips (some of which can be accessed via a network), designers can move from paper-and-pencil concepts to prototype hardware more easily. Many of the new high-performance computers on the market use processor chips custom-designed for that specific machine; automated chip design and manufacture shorten the time and improve the flexibility in producing custom chips.

Second, the market offers a variety of inexpensive, off-the-shelf chips that can be assembled to create new and interesting experimental designs. One of the best known successful examples of this type of research is a project initiated at the California Institute of Technology. There, researchers designed and built a customized computer to help them with certain specialized physics calculations. They developed the first ‘‘hypercube’ machine using a standard line of processor chips from Intel. Intel supported the project in the early days, principally through the donation of chips. Later, as the design concept proved itself and attracted the attention of government agencies, full-scale research support was provided to the group.

The impact of that low-budget project has been enormous. Several companies (including Intel) are in, or are planning to enter, the high-performance computer market with computers based on the hypercube design or one of its variations. Universities are beginning to realize the potential of specialized, low-budget machines, among them Caltech, Rice, and Syracuse. Three NSF centers (National Center for Supercomputing Applications, Pittsburgh Supercomputing Center, and the San Diego Supercomputer Center) also have installed these architectures for access by the nationwide academic community.

Based on the history and trends in computer architecture research, it appears that: 1) it is feasible to design and build computers with architectures customized for particular tasks; 2) the availability of powerful, inexpensive chips, has prompted academic laboratories to return to research in computer architecture; 3) new ideas in computer architecture can likely be commercialized quickly; and 4) universities that have access to fabrication facilities are more likely to develop new, specialized machines.

In the past, such customized machines would have been considered curiosities, with no chance of competing with traditional designs. The computer industry at that time was conservative, and users were unwilling to take chances on new ideas. Now, some entrepreneurs will gamble that if the system has distinct advantages in power and cost, new markets will open, even for systems based on radical new design theories.

But bringing a new high-performance machine to market is neither cheap nor simple. Millions of dollars-sometimes hundreds of millions-must be spent refining the design, developing software, and solving manufacturing problems, before a design concept moves from the laboratory into general use. The speed and ease of this transfer depends heavily on whether the technology is evolutionary or revolutionary.

It is difficult to say which computer technologies will become the foundation for building computers over the next decade. Despite the fact that all of the alternative technologies have difficulties to be overcome, it is likely that one or more new component technologies will be developed to fuel the rapid growth of computer capability into the next decade and beyond. But advances in fundamental technology alone will not be sufficient to achieve the increases in computer power that are needed by research users.

### Computer Architecture

The term “computer architecture” denotes the structural design of a computer system. It includes the logical behavior of major components of the computer, the instructions it executes, and how the information flows through and among those components. A principal goal of computer architecture is to design machines that are faster and more efficient for specific tasks.

‘‘Supercomputer’ is commonly used by the popular media to describe certain types of computer architectures that are, in some sense, the most powerful available. It is not, however, a useful term for policy purposes. First, the definition of computer ‘‘power’ is inexact and depends on many factors, including processor speed and memory size. Second, there is no clear lower boundary of ‘supercomputer power. IBM 3090 computers come in a wide range of configurations, but are they ‘‘supercomputers’ Finally, technology is changing rapidly, and with it the conceptions of the power and capability of various computers. Here, the term ‘‘ high-performance computers (HPC) (distinguished from the Federal program to advance high-performance computing referred to as the ‘‘high-performance computing initiative’ includes a variety of machine types.

One class of high-performance computing consists of large, advanced, expensive, powerful machines, designed principally to address massive computational science problems. These computers are the ones often referred to as "supercomputers." Their performance is based on central processing unit (CPU) power and memory size. They use the largest, fastest, most costly memories. A leading edge 'supercomputer’ can cost up to $20 million or more.

A large-scale computer’s power comes from a combination of very high-speed electronic components and specialized architecture. Most machines use a combination of “vector processing” and “parallel processing” (parallelism) in their design. A vector processor is an arithmetic unit of the computer that produces a series of similar calculations in an overlapping, assembly-line fashion (many scientific calculations can be set up in this way).

Parallel processing is the use of several processors that simultaneously solve portions of a problem that can be broken into independent pieces for computing on separate processors. Currently, large, mainframe high-performance computers such as those of Cray and IBM are moderately parallel, having from two to eight processors.[2] The trend is toward more parallel processors on these large systems. The main problem to date has been to figure out how problems can be setup to take advantage of the potential speed advantage of larger-scale parallelism.

The availability of software for supercomputer application is a major challenge for high-performance computing in general, but it is particularly troublesome in the case of large parallel processing systems. Parallel processing requires that the complexity of the problem be segregated into pieces that can run separately and independently on individual processors. This requires that programmers approach solutions in a very different manner from the way they program information flow and computations on vector processors. Until the art of parallel programming catches up with the speed and sophistication of hardware design, the considerable power of parallel computing will be underutilized. Software development for supercomputing must be given high priority in any high-performance computing initiative.

[2] To distinguish between this modest level and the larger scale parallelism found on some more experimental machines, some experts refer to this limited parallelism as “multiprocessing.”

Some machines now on the market (called minisupers' or 'minisupercomputers are based on the structure and logic of a large supercomputer, but use cheaper, slower electronic components and lower performance technology. They are relatively less expensive than high-end supercomputers. These systems sacrifice some speed, but cost much less to manufacture. An application that is demanding but does not require a full-size supercomputer may be more efficiently run on a minisuper.

Other types of specialized systems also have appeared on the market. These machines gain computation speed by using fundamentally different architectures. They are known by colorful names such as "Hypercubes,' "Connection Machines, " ' 'Data Flow Processors, " "Butterfly Machines," "Neural Nets, " or ''Fuzzy Logic Computers. " Although they differ in design concept, many of these systems are based on large-scale parallelism. Their designers get increased processing speed by linking large numbers-hundreds or even thousands— of simpler, slower, and cheaper processors. But computational mathematicians and scientists have not yet developed a good theoretical or experimental framework for understanding how to arrange applications to take full advantage of these massively parallel systems. Therefore, these systems are still, by and large, experimental, even though some are on the market and some users have developed applications software for them. Experimental as these systems are however, many experts believe that any significantly large increase in computational power must grow out of experimental systems such as these or from other forms of massively parallel architecture or hybrid architectures.

''Workstations, the descendants of personal desktop computers, are increasing in power; new chips being developed will soon offer computing power nearly equivalent to a Cray 1 supercomputer of the late 1970s. Thus, although high-end high-performance computers will be correspondingly more powerful, scientists who wish to do heavy-duty computing will have a wide selection of options in the future. Policy makers must recognize that:

- The term ''supercomputer' is a fluid one, potentially covering a wide variety of machine types; similarly, the '' supercomputer industry is increasingly difficult to identify as a distinct entity.
- Scientists need access to a wide range of high-performance computers from desktop workstations to full-scale supercomputers, and they need to move smoothly and seamlessly among these machines as their research needs require.
- Government policies should be flexible and broadly based to avoid focusing on a narrowIy defined class of machines.

Mere computational power is not always the sole objective of designers. For example, in the case of desktop computers like the Apple Macintosh or NEXT Computers, or the more powerful engineering workstations, much effort has gone into improving the communication between the machine and the operator (user interface). Computers are being designed to be more easily linked through data communication networks. Machines are being designed to do specialized tasks within computer networks, such as file management and internetwork communication. As computer designers develop a wider variety of machines specialized for particular tasks, the term ''high performance' covers a wider range of applications and architectures, including machines that are oriented to numerical scientific calculation.

## Computer Performance

Computers are often compared on the basis of computer power—usually equated to processing speed. The convention used for measuring computer power is "FLOPS" (floating point operations per second). The term ''floating point' refers to a particular format for numbers (scientific notation) within the computer that is used for scientific calculation. A floating point ''operation' refers to a single arithmetic step, such as multiplying or dividing two numbers, using the floating point format. Thus, FLOPS measure the speed of the arithmetic processor. Currently, the largest supercomputers have processing speeds ranging up to several billion FLOPS. DARPA has announced a goal of developing in this decade a ''teraflop' machine, a computer that executes one trillion FLOPS.

Peak computer speed and computer systems performance are two different things. Peak computer speed is the raw theoretical performance that is the maximum possible for the computer architecture. Computer system performance, the actual speed under use, is always lower—sometimes much lower. Theoretical peak speed alone is not a useful measure of the relative power of computers. To understand why, consider the following analogy.

At a supermarket checkout counter, the calculation speed of the cash register does not, by itself, determine how fast customers can checkout. Checkout speed is also affected by the speed that the clerk can enter each purchase into the cash register and the time it takes to complete a transaction with each customer—bag the groceries, collect money, make change—and move onto the next. The length of time the customer must wait in line to reach the clerk may be the most important factor of all, and that depends on how many clerks and cash registers are provided.

Similarly, in a computer, how quickly calculations can be set up and input to the processor and how quickly new jobs and their data can be moved in, completed, and the results moved out of the computer determines how much of the processor's speed can actually be harnessed (some users refer to this as ' 'solution speed"). Solution speed is determined by a variety of architectural factors located throughout the computer system as well as the interplay between hardware and software. Similar to the store checkout, as a fast machine becomes busy, users may have to wait in line. From a user's perspective, then, a theoretically fast computer can still deliver solutions slowly.

To test a machine's speed, experts use "benchmark programs, ' i.e., sample programs that reproduce a' 'standard' workload. Since workloads vary, there are several different benchmark programs, and they are continually being refined and revised. Measuring a supercomputer's speed is a complex and important area of research. Performance measurement provides information on what type of computer is best for particular applications; such measurements can also show where bottlenecks occur and, hence, where hardware and software improvements should be made.

One can draw some important implications from these observations on computing speed:

- Computer designers depend on feedback from users who are pushing their machines to the limit, because improvements in overall speed are closely linked to how the machines are programmed and used.
- There is no "fastest" machine. The speed of a high-performance computer depends on the skill of those that use and program it, and the type of jobs it performs.
- One should be skeptical of claims of peak speeds until machines have been tested by users for overall systems performance.
- Federal R&D programs for improving high-performance computing must stress software, algorithms, and computational mathematics as well as research on machine architecture.