(58)

FUTURE HIGH PERFORMANCE SCIENTIFIC COMPUTERS

Parallelism Is Now Creating Greater Potential Power; Are Users and Computer Specialists Ready?

by Gordon Bell, Vice President for Research and Development, Ardent Computer, Sunnyvale, California

The technology needed to bring about rapid increases in computational power is now available. The computers are being produced. Are enough people adequately prepared to take full advantage of these growing computational capabilities? I doubt it.

Parallelism makes the difference. Many new high-performance computers are now emerging for scientific and engineering applications and other high-speed, real-time uses. In addition to the more familiar gains due to improved VLSI (very large scale integration), their increased potential is based upon parallelism.

The Strategic Computing Initiative (SCI) sponsored by the Defense Advanced Research Projects Agency, is stimulating mainline manufacturers and new companies, backed by venture capital, to build these new computers, in conjunction with the university research community. The research pipeline and the mechanism for transferring innovations from the campus to the marketplace are clearly working.

Furthermore, a recent report from the White House Office of Science and Technology Policy has recommended an initiative for high-performance computing. It is to include increased funding for new components and systems, training, and greater emphasis on research responding to the "grand challenges," based upon computational science, and a National Research Network.

Gains in performance due to improvements in circuit hardware have been taking place at a rate of14% per year—roughly a doubling of performance every five years. Today, greater power from a single computer system due to parallelism is finally increasing more rapidly than gains due to circuit improvements. Increasing the numbers of parallel processors can bring about performance gains of a factor of two every two years: a factor of 32 in a decade.

This maximum potential power will not be available to a single job unless parallelism is tamed or users are retrained. Neither the user community nor the computer science community is moving rapidly enough to understand and exploit these potential gains in performance.

For the first time in the history of computations, virtually all new high-performance computers—for example, the multiprocessor product lines from Cray Research and ETA—provide a high degree of scalability. Each product series is based upon a common root processor, hardware components, and operating system. With scalability, it is possible to produce a number of different computer systems, of various sizes and corresponding levels of performance, that have basically the same architecture.

For the next generation (1988-1994), the main line of high performance computers made in the United States will form an evolving range that, from today's perspective, will look like this:

- Supercomputers: Priced at $10 million or more, from Cray Research, ETA, and IBM;

- Minisupercomputers: $1 million or less, from Alliant, Convex, Cydrome, Floating Point Systems, and Multiflow;

- Graphics supers: Under $100 thousand, from Ardent and other manufacturers, which provide traditional supercomputer style access coupled with very high speed interactive graphics for "visualizing" the computation;

- Micro-supers: Attachable to a personal computer, in the $10 thousand range; not yet in the market, but to be expected.

The summit of this traditional range will be occupied by the CRAY-4, targeted for 1992. It will use 64 processors, each capable of two billion floating point operations per second (gigaflops), for a total potential performance rate of 128 gigaflops.

No economy of scale, measured in terms of processing operations per second per dollar, will be observable over the range of personal supers, graphics supers, minisupers, and full-scale supers. The new class of graphics supers offers readily-available, close-at-hand power and exceptional visual capabilities to the user, but these machines appear to provide a diseconomy of scale for general-purpose computing. The micro-super or personal supers may well offer the best performance/cost ratio. The reason for the diseconomy of scale is that VLSI CMOS technology is gaining speed and density more rapidly than bipolar (now LSI) technology. Fast machines pay extensive penalties for primary memory and disk technology because it does not scale in either cost per bit or performance.

Plain Old one-chip microprocessors (POP's) are becoming very fast. For scalar/integer work, they approach the speed of the largest mainframes and supercomputers. Attached vector units will make such uniprocessors very useful and cost-effective in workstations and small computers.

A switch from CMOS technology to bipolar VLSI technology could make it possible to achieve an increase in performance of a factor of five to ten by the early 1990's. Today's reduced instruction set computers (RISC), using fast single-chip microprocessors based upon a smaller instruction set and extensive pipelining, do logical and integer operations at nearly the speed of supercomputers.

Multicomputers are a collection of 32 to 1024 interconnected computers, each with its processor and memory, communicating with one another by passing messages. These are the most cost-effective for single scientific jobs, provided the problem is compatible with the computer.

Large-scale parallel machines without vector processing, using hundreds of microprocessors similar to those used on personal computers, are being proven to be cost-effective for selected scientific and engineering applications. However, users must rewrite their programs for the computer. By improving the relatively poor floating point performance of today's microprocessors and improving parallelizing compilers, these computers could turn into the mainstream within five years.

Today, large-scale multiprocessors are clearly superior to existing systems for transactions processing, batch processing, and program development. Furthermore, by using the same basic components, a wide range of computers from two to several hundred processors can be constructed. This achieves one of the greatest dynamic ranges of scalability.

The Institute of Electrical and Electronic Engineers (IEEE), recently awarded prizes to recognize work on parallelism. Noteworthy programs included:

- The climate model at the National Center for Atmospheric Research, which achieved over 450 megaflops on the CRAY X-MP/416, or over 50% of its four-processor peak vector processing capability;

- Sandia National Laboratory's programs for Beam Stress Analysis, Unstable Fluid Flow, and Baffled Surface Waves; these operated at speedups of 400 to 1020 times, compared with results on an individual processor, on the N-CUBE computer system composed of 1024 independent computers.

A single-instruction, massively large data (SIMD) computer, the Connection Machine, has been used as a supercomputer for several applications. The current Connection Machine model, CM2, is scalable and comes in a variety of sizes, ranging from 8192 to 65,536 processing elements. Like other applications-specific computers, the Connection Machine runs only one (or a few) programs at a given time. The price/performance advantage compared with a general-purpose computer ranges from a factor of ten to a factor of 100.

A variety of other computer designs based upon parallelism have proven themselves. They are either emerging into the market or show great promise. These are truly applications-specific and are used to carry out tasks including vision, speech, and text analysis and databases. The application is bound to the hardware as well as the software when the machine is designed and constructed. Users thus might obtain an improvement of a factor of 10,000 in performance, cost reduction, or performance/price.

The best example of such a computer is a specialized processor for speech processing. The total cost of components used to make the device was less than $100,000. It computes 32-bit floating-point numbers at the rate of one-quarter trillion per second, going beyond gigaflops to function at 0.25 teraflops.

The path now seems clear toward a computer capable of executing one trillion operations per second for a relatively large class of applications by 1994. Possible solutions include the multicomputer, such as the TF1 being constructed at an IBM research laboratory, composed of 32,768 fifty-megaflop computers, and the SIMD approach, such as the Connection Machine.

Computer science has yet to embrace vectors as a machine primitive to be incorporated in texts and courses. It is necessary to install, use, and understand both vector and parallel machines. Texts must be written, and students trained. An aggressive program to address all forms of parallelism is needed. The most serious training programs are at the supercomputing centers supported by the National Science Foundation. This is limited to the mundane problem of vectorizing Fortran programs on single processors.

I believe that the main barriers to progress in supercomputing based upon vector and parallel processing are:

- A lack of understanding of the potential of interactive displays to visualize results, especially when coupled to high performance;

- Insufficient and unsatisfactory training;

- Inadequate involvement by computer scientists and computational engineers in seeking to achieve high performance on real scientific and engineering applications.

- Lethargy on the part of users and a lack of coupling to computer science. ❐