How to Build a HYPER computer

BY THOMAS STERLING

The simulation and ultimate solution of humanity's major ills and most perplexing problems require significantly faster supercomputers

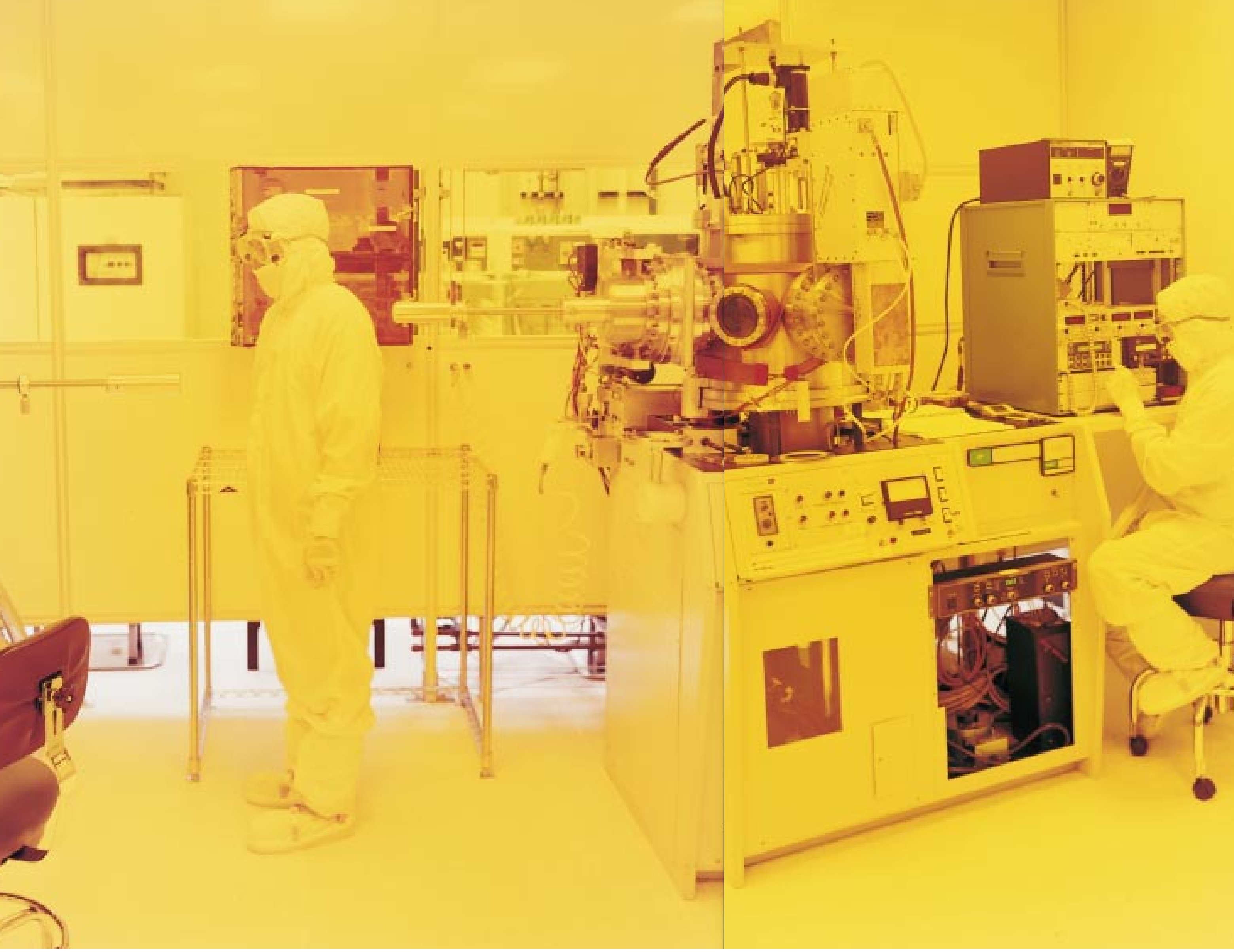
Photographs by Olivier Laude

TODAY'S FASTEST supercomputers run too slowly to do tomorrow's science. Despite the ongoing revolution in communications and information processing, many computational challenges critical to the future health, welfare, security and prosperity of humankind cannot be met by even the quickest computers. Crucial advances in pivotal fields such as climatology, medicine, bioscience, controlled fusion, national defense, nanotechnology, advanced engineering and commerce depend on the development of machines that will operate at speeds at least 1,000 times faster than today's biggest supercomputers [see "Crucial Tasks for Hypercomputers," on page 41].

Solutions to these incredibly complex problems hinge on the ability to simulate and model their behavior with a high degree of fidelity and reliability, often over long periods. This level of performance goes far beyond that of present-day supercomputers, which at best can execute several trillion floating-point operations per second (teraflops). It could take 100 years, for example, for the largest existing system to perform a complete protein-folding computation—a long-sought capability. To accomplish this kind of analysis task, researchers need hypercomputing systems that achieve at least petaflops speeds—that is, more than a quadrillion floating-point operations (arithmetical calculations) per second.

Not only do current high-end computers run too slowly, they cost too much. The three-teraflops (peak performance) ASCI (Accelerated Strategic Computer Initiative) Blue systems that are dedicated to the stewardship of the U.S. nuclear stockpile cost approximately \$120 million each. That's equivalent to a price/performance factor of \$40 per peak megaflops (million flops), which is more than 10 times greater than the price/performance of a premium personal computer. High-end computers impose indirect costs as well. Annual payments for

QUICK COMPUTER: The three-teraflops ASCI (Accelerated Strategic Computer Initiative) Blue system at Lawrence Livermore National Laboratory helps to maintain the nation's nuclear weapons stockpile.



FAST CHIPS: Engineers at TRW's Space Park facility in Redondo Beach, Calif., use sputtering machines to deposit thin superconducting films on silicon wafers as part of the fabrication of prototype superconducting processor chips.

the electrical power to operate such systems can easily exceed $1 million. Housing their oversize footprints can also add significant expense. Paying crack programmers to write the complex code for these machines is yet another cost.

Despite their impressive processing speeds, high-end systems do not make good use of the computing resources they have, resulting in surprisingly low efficiency levels. Twenty-five percent efficiency is not uncommon, and efficiencies have dropped as low as 1 percent when addressing certain applications.

The hybrid technology multithreaded (HTMT) system is a new class of computer that offers 100 times the capability of present high-end machines for roughly the same cost, power usage and floor space. Further development could bring the technology beyond a quadrillion flops to trans-petaflops territory—1,000 times the performance of today's best systems or more. To achieve these goals, a multi-institutional, interdisciplinary team has created a computer architecture able to harness various advanced processing, memory and communications technologies, leveraging their strengths and complementing their limitations. The basic elements of HTMT have been developed with financial support from NASA, the National Security Agency, the National Science Foundation and the Defense Advanced Research Projects Agency; actual construction awaits further governmental funding.

Ironically, it is the very success of computing technology that reveals its limitations. Back in the late 1970s, personal computers could barely play Pong. A system capable of executing a major science problem of the day at a performance level of a few tens of megaflops could cost $40 million or more. In contrast, PCs now priced at less than $2,000 can outperform those machines.

Historically, the supercomputer industry has pushed the frontiers of processing performance with a combination of advanced technology and architectures customized to address specific problems. The unfortunate side effect has been high price tags. Exorbitant costs and lengthy development times have kept the market for such systems relatively flat while other segments of the computer industry have grown explosively. With these costs forcing up the price to the customer, the overall supercomputer market and corporate investment in the technology have remained limited, producing a classic commercial death spiral.

Even when alternative approaches have been tried—including custom vector computer architectures (which efficiently perform a single operation on a list of numbers using pipelined memory access and arithmetic functional units) as well as massively parallel systems integrating large arrays of cooperating microprocessors—the costs of such systems have remained high while operational efficiencies for many applications have suffered. In the past two or three years, a number of groups have built highly parallel general-purpose computers with peak performance levels of more than a teraflops. Yet low efficiency levels mean that little of this processing capability can be brought to bear on real-world applications. As a result, commodity clusters—networked arrays of standard computing subsystems—are perceived as the only economically viable pathway: they require little additional development in spite of the programming difficulties and communications delays inherent in using clustered systems.

Research on new classes of petaflops-capable systems has been under way since the mid-1990s. Engineers have been attacking the speed problem on all fronts, pursuing various technology paths to such machines. With sufficient R&D support, all can be accomplished within this decade [see "Five Routes to Ultrafast Processing," on the next page]. Although each method has its strengths and weaknesses, one of the most widely applicable is the HTMT design.

HTMT exploits a diverse array of advanced technologies within a single flexible and optimized system. The project attempts to achieve efficient trans-petaflops performance by incorporating superfast processors, high-capacity communications links, high-density memory storage and other soon-to-mature technologies in a dynamic, adaptive architecture.

No matter what course they take, designers of trans-petaflops systems all face three challenges. First, they must find a way to aggregate sufficient processing, memory and communications resources to achieve the targeted peak-computing capabilities despite practical constraints of size, cost and power. The second goal is to attain reasonable operational effi-

Crucial Tasks for Hypercomputers

Many intricate scientific problems with enormous social and political implications await solutions that can be processed only on computers that can execute more than a quadrillion floating-point operations per second—trans-petaflops performance.

Climate Modeling

Perhaps the most critical issue facing the earth's inhabitants is the need for accurate predictive scenarios for both short- and long-term weather changes. First, trans-petaflops computers could integrate the huge quantities of satellite data into detailed maps. The mapped data could then be used to simulate and model the chaotic and interrelated behaviors of the elements of our global climate system, allowing accurate predictions.

Controlled Fusion

Both an answer to the world's energy problems and a way to power spacecraft across the solar system, thermonuclear fusion's vast complexity has kept it continually just over the horizon. Trans-petaflops computers would simulate the thermal, electromagnetic and nuclear interactions of large numbers of particles in a dynamic magnetic medium to help in designing practical fusion reactors.

Medicine/Bioscience

Considerably faster computing capability could give medicine the edge in combating continuously evolving diseases. This job requires molecular-level analysis to achieve nearly instantaneous drug design, including exploring complex protein folding.

Agriculture

To feed the earth's ever growing population, rapid computation will help develop new genetically engineered crops and solve the complex problems involved in managing the world's ecology.

National Defense

With the real-world testing of nuclear weapons banned, trans-petaflops machines could model the behavior of these systems to help maintain the readiness of the strategic weapons stockpile. Real-time decryption of increasingly complicated secret codes is one key to maintaining national defense.

Commerce and Finance

Large-scale mining of the enormous data spaces containing business information and economic statistics will allow a more accurate simulation of commercial systems.

Nanotechnology

With digital electronics shrinking to the atomic scale, where quantum mechanics is important, chip designers can no longer model electronics using averaged physical parameters.

Advanced Engineering

Ultrafast processing will be needed to simulate the behavior of new materials and composites at the microscale. Future aircraft design and that of other complex engineered systems will benefit from the same type of detailed modeling capabilities.

Astronomy

To model the galaxy and its 100 billion stars properly, new, superfast computers will be required to analyze the complex interplay of the interstellar medium and heavier molecules.

ciencies in the face of standard degradation factors. These include latencies (time delays) across the system, contention for shared resources such as common memory and communications channels, overhead-related resource reductions caused by the need to manage and coordinate concurrent tasks and parallel resources, and wastage of computing resources (starvation) caused by insufficient task parallelism or inadequate load balancing. The third objective concerns finding ways to improve the usability of the system—a somewhat arbitrary measure comprising the issues of generality (general utility), programmability and availability.

Superconducting Processors

DURING THE PAST DECADE, digital logic has been dominated by CMOS (complementary metal oxide semiconductor) processors. CMOS technology has provided lower power and greater performance while system densities have increased at an exponential rate. Yet the fastest digital logic technology on earth is not CMOS. An altogether different technology using another kind of physics claims that title: superconducting logic.

Discovered at the beginning of the 20th century, superconductivity is the ability to conduct electricity with no resistance, a phenomenon that some materials exhibit when cooled to cryogenic temperatures. In principle, a loop of superconducting wire can sustain an electric current forever. More important, superconducting devices exhibit quantum-mechanical behavior in macroscale electronic components and circuits. In the early 1960s researchers developed a nonlinear switching device based on superconductivity called the Josephson junction, which was found to have exceptional speeds.

The HTMT hypercomputer design will employ high-speed superconducting logic processors based on Josephson junction technology. In rapid single-flux quantum (RSFQ) technology, supercon-

Five Routes to Ultrafast Processing

One approach to attaining trans-petaflops computing performance (more than a quadrillion floating-point operations per second) is to use a hybrid architecture combining several soon-to-be-available advanced technologies (*see accompanying article*). Here are five other technical pathways to achieving that goal.

NAME	METHOD	EXAMPLE	BEST APPLICATIONS
1 SPECIAL-PURPOSE ARCHITECTURE OR SYSTOLIC ARRAY	Specially designed hardware and software that mirror the abstract problem to be solved. Runs in parallel with a fast data pipeline to speed computation	Grape Project (University of Tokyo)	Huge multibody calculations, stellar cluster simulation, bioinformatics
2 CELLULAR AUTOMATA	Finite-state machine in which many relatively simple computing cells in a large 2-D or 3-D matrix operate in lockstep during each clock cycle. Each cell's actions depend on its internal state and those of its nearest neighbors	Never fully executed	Computational fluid dynamics, diffusion simulations
3 PROCESSOR-IN-MEMORY (PIM) ARCHITECTURE	With a good deal of processing and memory on each chip, the system logic sees all the bits coming out of the dynamic random-access memory (DRAM) at the same time. There's lots of memory access and little delay in data transmission speed processing during each cycle	IRAM (University of California at Berkeley) Blue Gene (IBM)	Image processing, data encryption, rapid database searches, protein-folding modeling
4 BEOWULF OR CLUSTER ARCHITECTURES	A high-bandwidth mesh system interconnects many low-cost, commodity processors (each a partial-system-on-a-chip device) in a high-density array	GigAssembler software (International Human Genome Sequencing Consortium)	Wide range of problems; deciphering the human genome
5 DISTRIBUTED COMPUTING OR MEGACOMPUTING ARCHITECTURES	Harness the unused computing cycles on the estimated 500 million personal computers linked to the Internet. Inefficient communications is a drawback	SETI@home (Serendip Project)	Huge parallel problems such as Monte Carlo simulations and monitoring the function of the Internet

Not only do current HIGH-END SUPERCOMPUTERS run too slowly, they COST TOO MUCH and use too much power.

ducting loops store information as tiny magnetic flux quanta (by discrete current levels). The loops, called superconducting quantum interference devices, or SQUIDs, are simple mechanisms originally developed as sensing devices that comprise two Josephson junctions connected by an inductor, which is like a solenoid. With both Josephson junctions operating, a current injected into the loop will continue indefinitely. SQUIDs exhibit the interesting characteristic of having distinct states of operation: they may contain no current, sustain the basic current, or have a current that is some integral multiple times the basic current but nothing in between. This remarkable property results from quantum-mechanical effects. To represent the 0's and 1's of digital code, RSFQ logic gates use discrete currents (or fluxes) rather than distinct voltage levels. When cooled to a temperature of four kelvins, these units can operate at more than 770 gigahertz, the fastest (single-gate) processing speeds ever achieved and approximately 100 times quicker than conventional CMOS logic.

RSFQ technology will allow the hybrid computing system to run nominally at from 100 to 200 gigaflops (billion flops) per processor as opposed to a few gigaflops, as in standard CMOS processors. In addition, the minuscule and packetized nature of magnetic flux quanta in RSFQ devices cuts crosstalk and power consumption by a couple of orders of magnitude. This rapidly maturing technology reduces parallelism requirements, cost, power demand and system size.

Boosting Efficiency

WITH SUPERFAST processors in place, HTMT seeks to make efficient use of their powerful capabilities. Those processors should spend their time doing little else but computations. Conventional approaches such as commodity clusters require large-scale tasks to be run on similarly large-scale computational nodes. Often a computational node on a conventional system must wait while a remote request to another node is being serviced. Unless operators exactly balance the workload, some nodes will continue to compute while others, having finished their jobs, will stall. Even when engineers employ load-balancing software techniques, the overhead required for accomplishing this function can reduce efficiency.

Unlike any other computer architecture, HTMT revolutionizes the relation between the processing system and the memory system. In ordinary multiprocessor systems, the computational processors manage and manipulate the "dumb" memory system; in contrast, HTMT's "smart" memory system administers the processors. HTMT and other tightly coupled parallel computers consider the workload on the processing elements and make on-the-fly decisions as to which part of a task should be performed by what hardware. In doing so, the processors work out of their local registers and some high-speed buffer memories, thus avoiding having to reach too far out into the system. The result is a drop in latency problems. The processors do not spend time managing memory resources, which are just wasted processing cycles that add to overhead; these logistical decisions are made by the small low-cost processors in the memory.

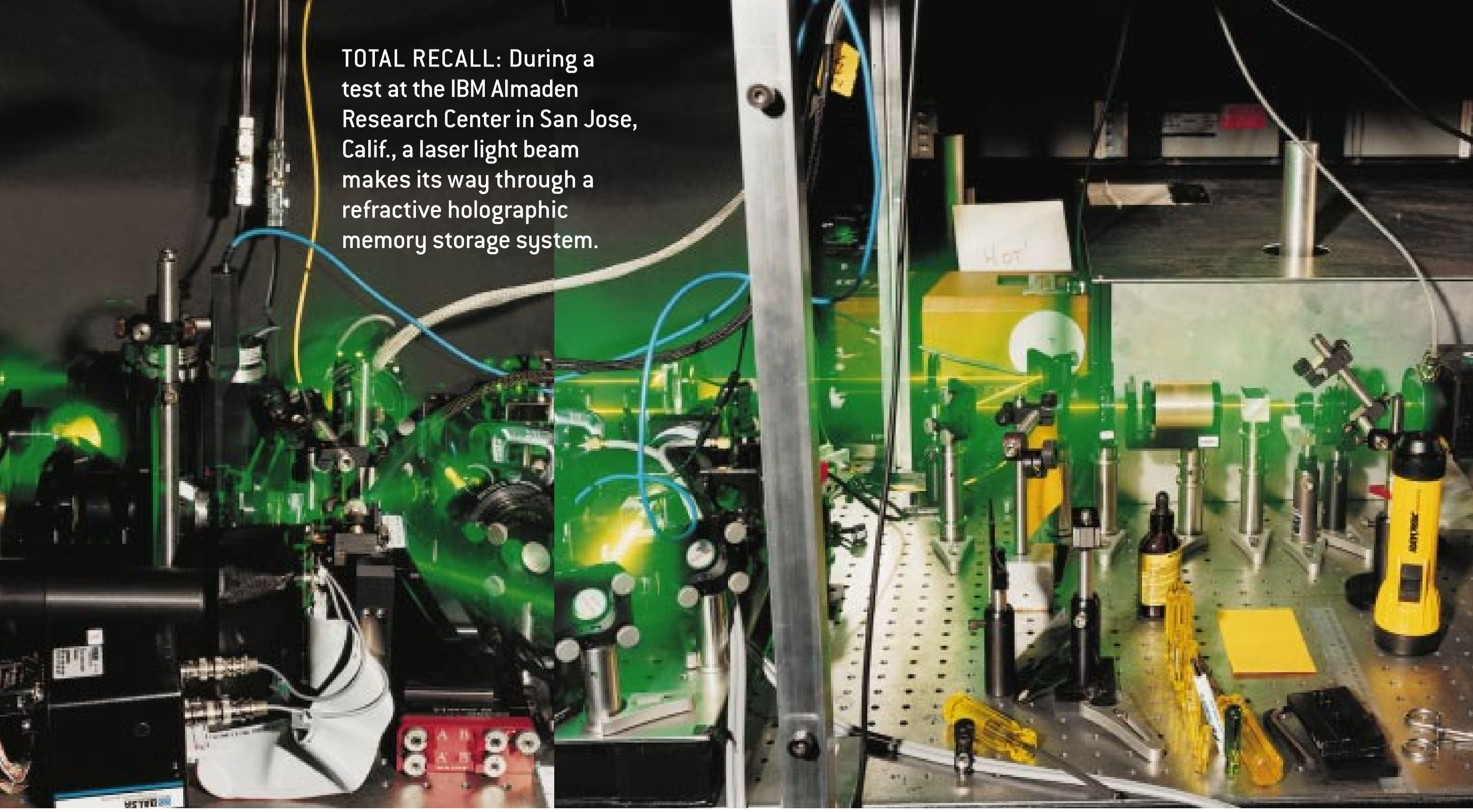
The HTMT design attacks the problem of latency in two ways. First, the system employs a dynamic, adaptive resource management scheme based on a multithreaded architecture that enables HTMT to switch from one stream of instructions to another within a single cycle. Whereas most computers operate with one stream of instructions, HTMT will feature multiple instruction streams. By using overlapping communications, the processors can work on many outstanding requests simultaneously. Say a superconducting processor needs to load information from a cache or a high-speed buffer, a procedure that will take many 10-picosecond cycles. As this request is served by the memory system, the processor can switch to another data stream to find operations that can be performed immediately.

The second way HTMT will handle the latency issue is by employing processor-in-memory technology (PIM), wherein small secondary satellite or taxi logic processors are placed in its memory devices. A few years ago fabrication advances allowed CMOS logic and dynamic random-access memory (DRAM) cells to be put onto the same silicon die, permitting them to be closely integrated. These cheap devices deal with overhead—that is, they manipulate information in the memory, again allowing the superconducting processors to focus on computing. PIM processing technology can also handle memory-intensive functions such as data gathers—collecting needed information from various locations and placing it in one dense object—as well as carrying out the reverse operation of data scatters—distributing the information to the correct locations.

Although the technologies and architecture incorporated by the HTMT system may be innovative, the means of managing these resources and the computing discipline employed by HTMT are truly revolutionary. The system will use new percolation techniques in which the PIM processors decide when a new piece of work should be performed. They will determine when to migrate all information that needs to be executed up to rapid-access buffer memories near the high-speed superconducting processors.

THE AUTHOR

THOMAS STERLING holds a joint appointment at the NASA Jet Propulsion Laboratory's High Performance Computing group, where he is a principal scientist, and the California Institute of Technology's Center for Advanced Computing Research, where he is a faculty associate. For the past 20 years, Sterling has carried out research on parallel-processing hardware and software systems for high-performance computing. Since 1994 he has been a leader in the national petaflops initiative. He heads the hybrid technology multithreaded architecture research project.



TOTAL RECALL: During a test at the IBM Almaden Research Center in San Jose, Calif., a laser light beam makes its way through a refractive holographic memory storage system.

For example, when a specific subroutine is required, it and the special information it needs to execute its function will be moved up to the processors. This proactive method of prestaging necessary information is a way to avoid creating long latency delays in connecting to the main memory. The technique also frees the high-speed processors from having to perform logistical overhead operations, because they are not needed to bring information to the processing sites.

Improving Usability

THE THIRD MAJOR challenge to transpetaflops computing concerns the usability of the system: researchers must increase its generality (to ensure that it can handle a wide variety of problems), make it easier to program, and boost its availability, or uptime. HTMT addresses these issues in several ways.

By using a global name space in a shared-memory computing structure, every processor can "see" all of the memory. This method is more general than typical distributed- (or fragmented-) memory computing techniques because it provides efficient access by any processor to all data without having to engage software routines on a remote processor to assist in the data transfer. More actions can be performed simultaneously, speeding execution. In addition, by letting the system conduct dynamic rescheduling—responding to run-time information—it can perform certain computations more effectively, a capability that adds to its generality. And because this arrangement is closer to the way computational scientists think about their problems, programming the system is more intuitive. Typically programmers must determine beforehand how a problem should be handled by a system, a complex and laborious task. But an HTMT system makes many of these decisions by itself, thereby helping to alleviate one of the biggest difficulties in working with large computers—programming them.

The hybrid computing system will

Key Concepts/*Hypercomputing*

Contention—Time delay created when two processors try to access a shared resource simultaneously

Latency—Delay caused by the time it takes for a remote request to be serviced or for a message to travel between two processing nodes

Load Balancing—Distributing work evenly so that all processing nodes are kept occupied as the program is executed

Overhead—Time spent on noncomputational functions such as the logistical management of parallel resources and concurrent tasks

Percolation—Method of managing tasks and data movement without incurring delays caused by overhead, latency, contention or starvation

Processor-in-Memory (PIM)—Integrated circuits that contain both memory and logic on the same chip

Starvation—Wastage of computing resources caused by insufficient program parallelism or poor load balancing

Wave Division Multiplexing (WDM)—Method by which the effective bandwidth of an optical channel can be increased by using optical signals with different wavelengths

The hybrid computer REVOLUTIONIZES the relation between the PROCESSING system and the MEMORY system.

provide greater availability to users through the use of higher-capability subcomponents, allowing it to achieve the same level of performance with fewer parts. This parts reduction increases the mean time between failures of the entire system, thus boosting operational uptime.

Holographic Memory Storage

ANOTHER INNOVATIVE aspect of the HTMT system will be its use of high-density-capacity holographic memory storage devices. This alternative to the semiconductor-based DRAM is being explored by academic and industrial research laboratories and should provide superior storage density as well as lower power consumption and costs.

Holographic storage systems use light-sensitive materials to accumulate large blocks of data. Photorefractive and spectral hole–burning techniques represent two distinct approaches. In photorefractive storage, a plane of data modulates a laser beam (signal) that interferes with a reference beam in a small rectangular block of a storage material such as lithium niobate. The hologram results from the electro-optic effect that occurs when local electric fields are created by trapped, spatially distributed charge carriers excited by the interfering beams. Many data blocks may be stored in the same target material. They are differentiated by varying either the angle of incidence or the wavelength of the laser beam. The spectral hole–burning technique relies on a nonlinear response of a storage material to optical stimuli. Data are represented by changes in the photosensitive medium's absorption spectrum. Many bits can be stored at a given spatial location.

Photorefractive methods are more far advanced. But in the long-term, spectral hole–burning technology may yield significantly higher memory density. Typical holographic devices currently feature access times of several milliseconds—approximately the same as conventional secondary storage devices such as hard disks and CD-ROM drives. But advanced techniques employing tunable lasers or arrays of laser diodes each set at a slightly different angle to one another are expected to yield access times of a few tens of microseconds. Although these access times are about two orders of magnitude longer than that of DRAM, their data bandwidths are the same or greater, and the systems are about 100 times faster than conventional disk drives. Storage capacities of 10 gigabits or more in blocks as small as a few cubic centimeters are expected within the next decade.

Optical Communications

TO CONNECT THE SPEEDY superconducting processors and high-density holographic memory systems in a network, HTMT will use high-capacity optical data pipelines. Instead of employing electrons in metal wires, HTMT will speed communications by using photons in fiber-optic cables. Wires can easily handle hundreds of megabits per second, and speeds of a few gigabits per second (gbps) can be achieved by using differential pairs of input/output pins (one goes up while the other goes down). But it could take tens of millions of wires to supply all the global communications bandwidth required of systems operating in the petaflops regime. With modulated lasers, digital light signals can transmit at up to 10 gbps per channel or more in conventional optical communications systems.

Employing multiple wavelengths (or colors) of light carrying digital information dramatically improves fiber-optic bandwidth or channel capacity. HTMT will use an advanced optical transmission system called wave division multiplex (WDM) communications. It should provide about 100 times the per-channel bandwidth of the best conventional metal-wire communications systems. WDM allows separate digital signals, each with its own dedicated light wavelength, to travel together through the same channel. The number of different wavelengths that can be simultaneously transmitted through a single channel has grown to around 100 in recent years, and in time this figure may rise further. With improved receiver, transmitter and switch technology now in development, switching rates of 50 megahertz or more will soon be possible. Still experimental devices may bring about rates on the order of one gigahertz in the future. This capacity level would be sufficient to manage the huge information flow of a petaflops-scale computing system.

These next-generation hypercomputers would offer an important tool for exploring the world's most pressing problems, including global warming, disease epidemics and cleaner energy. In 1999 the President's Information Technology Advisory Committee strongly recommended financial support for these kinds of projects. Research groups have demonstrated that HTMT technologies could be the best route to trans-petaflops performance. Proper funding is all that is needed to put these systems in place. SA

This article is the first in a two-part series on next-generation supercomputers. The second part, "The Do-It-Yourself Supercomputer," will appear in the August issue.

MORE TO EXPLORE

Challenges of Future High-End Computing. David H. Bailey in *High Performance Computer Systems and Applications.* Edited by Jonathan Schaeffer. Kluwer Academic Publishers, 1998. Preprint available at **www.nersc.gov/~dhbailey/dhbpapers/future.pdf**

In Pursuit of a Quadrillion Operations per Second. Thomas Sterling in *NASA HPCC Insights,* No. 5; April 1998. Available at **www.hpcc.nasa.gov/insights/vol5/petaflop.htm**

The author's Web site: **www.cacr.caltech.edu/~tron/**

A Hybrid Technology Multithreaded (HTMT) Computer Architecture for Petaflops Computing. Thomas Sterling. On the JPL/NASA Project HTMT Web site at **http://htmt.jpl.nasa.gov/intro.html**

NASA high-performance computing and communications Web site: **www.hq.nasa.gov/hpcc/petaflops/**