innovations

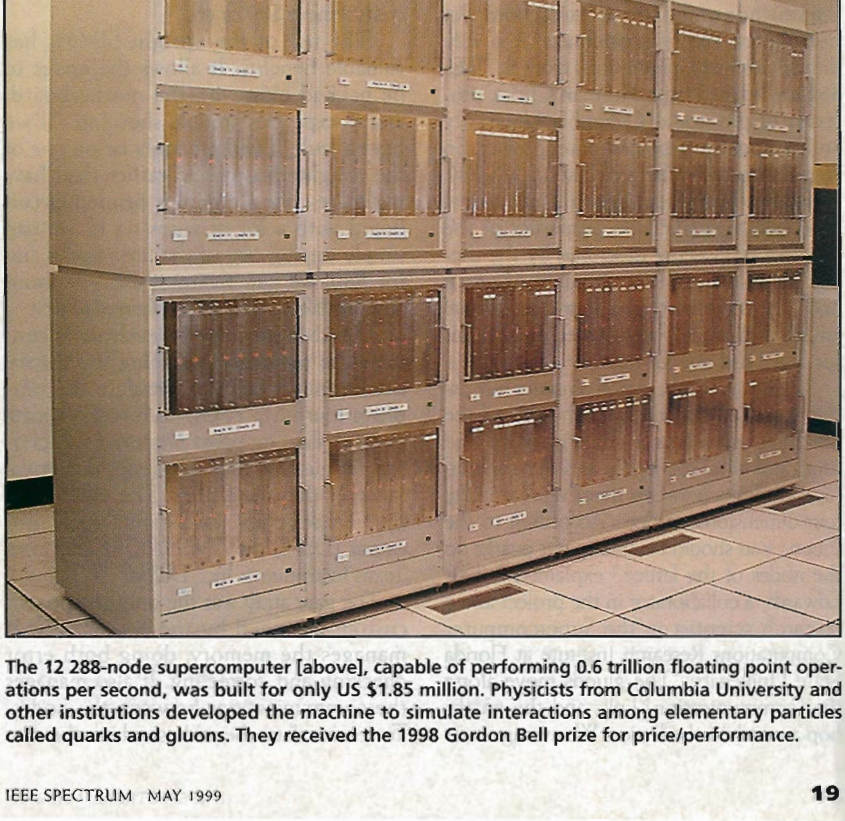
Champagne supercomputer on a beer budget

Every year, at the High Performance Networking and Computing Conference, Gordon Bell prizes are conferred for highly parallel programs and systems. One award honors the best achievement in high-performance computing. The second goes to the supercomputer with the best price/performance ratio. The prizes' sponsor is computer pioneer Gordon Bell, a senior researcher at Microsoft Corp., Redmond, Wash.

In 1998, the winner of the price/performance prize was a team of physicists from Columbia University in New York City plus their collaborators from four other institutions. The machine they constructed contains 12 288 processor nodes and was assembled for a mere US $1.85 million [see photo]. Yet it simulates the behavior of elementary particles and is designed to explore the first picoseconds of the universe's existence.

Housed at the Riken Brookhaven Research Center in Upton, N.Y., the supercomputer has a peak performance of 0.6 trillion floating-point operations per second (Tflops). (By way of comparison, one of the fastest machines in the world, a supercomputer at Sandia National Laboratories in Albuquerque, N.M., is capable of three times as many floating-point operations per second, but at $55 million, has cost nearly 30 times as much to build!)

Several similar but smaller machines constructed by the team are located around the globe. They include a 64-node machine at the University of Wuppertal in Germany; a 128-node machine at Ohio State University, Columbus; one with 1024 nodes at Florida State University in Tallahassee; and a 8192-node machine at Columbia University in New York City.



The 12 288-node supercomputer [above], capable of performing 0.6 trillion floating point operations per second, was built for only US $1.85 million. Physicists from Columbia University and other institutions developed the machine to simulate interactions among elementary particles called quarks and gluons. They received the 1998 Gordon Bell prize for price/performance.

The 'quarks' behind the machine

The structure of the machine reflects the nonlinear physics it was designed to simulate. The fundamental theory of strong interactions, or quantum chromodynamics, deals with elementary particles called quarks. There are six types of quarks in all, but two of them—the up- and the down-quark—form the protons and neutrons that form the nuclei of ordinary matter.

Two up-quarks and a down-quark bind together into a proton. Two down- and an up-quark make a neutron. The up-quark carries a positive charge that is two-thirds the value of an electron's charge. The down-quark carries a negative charge one-third of an electron's charge.

For each quark there is a so-called anti-quark, which has the same mass as the quark but a charge of opposite polarity. Einstein's theory of relativity states that under the right conditions, such particle-antiparticle pairs can be created and destroyed.

What binds the quarks together is another elementary particle called, appropriately enough, a gluon. In some ways, quantum chromodynamics is similar to quantum electrodynamics. In the latter theory, charged particles—electrons and protons, for example—interact through the exchange of photons. But while photons do not interact with one another, gluons do. This is what makes quantum chromodynamics a nonlinear theory.

"Since the early '70s, physicists have believed that they understood the underlying structure of the strong interactions. But because the theory is nonlinear, it is very difficult to make predictions from it," *IEEE Spectrum* learned from Norman Christ, professor of physics at Columbia University and head of the supercomputer project. So rather than solve the problem analytically, physicists have been using a numerical approach to predict the properties of the particles supposedly described by the theory. In essence, the continuum of space and time is replaced by a four-dimensional lattice of discrete points: three dimensions for space and one for time.

The design of the supercomputer mirrors this numerical approach. Each node of the machine represents a point in space and time, and all of the nodes and the communication links between them represent the four-dimensional lattice. "According to the theory, you should be able to put quarks on the nodes of the lattice," explained Robert Edwards, a collaborator in the project and a research scientist in the Supercomputer Computations Research Institute at Florida State University. "The gluons move along the [communication] links and the quarks hop around on the lattice. We average over all the possible ways you could put the quarks on the lattice and the relative weights of how strongly they interact with one another through the links."

This method enables the scientists to calculate many properties of interest, such as particle masses and decay rates. The number of quarks on each of the nodes is not predetermined, though. "If the temperature and interaction strength are high enough, particle-antiparticle pairs will be created and destroyed," said Christ. "As the calculation proceeds, the number of particles fluctuates in a way that we do not specifically control."

Another aspect of physical theory that bears on the computer design is that all interactions are local. Consequently, information cannot be sent to a particular node without first being passed along from neighbor to neighbor in the lattice until it arrives at its destination. In other words, each node of the computer communicates with only its eight nearest neighbors: six neighbors in space and two—backwards and forwards—in time. The direction of data flow in two dimensions is illustrated in the figure [bottom], where the blue dots represent the nodes and the red arrows indicate the active links and the direction of data flow.

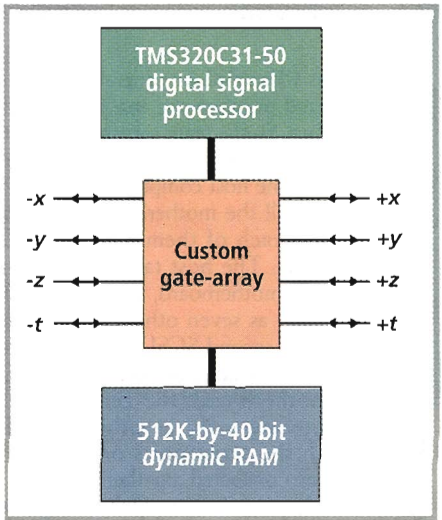
Replaceable nodes

Each node of the machine occupies a daughterboard, measuring about 45 mm by 68 mm and holding a processor, 2MB of dynamic RAM, and a custom-designed VLSI gate array chip mounted on a single in-line memory module (SIMM). All this costs a mere $80 or so.

The use of the upright SIMMs has valuable benefits. It allows the nodes to be packed closer on the motherboard, and it expedites trouble-shooting. "If we have a fault, it will probably be on one of the daughterboards. So rather than having to rework a complex printed-circuit board with a lot of wires on it, we can simply unplug one of the little nodes and for $80, we can throw it away without any terrible hardship," explained Christ.

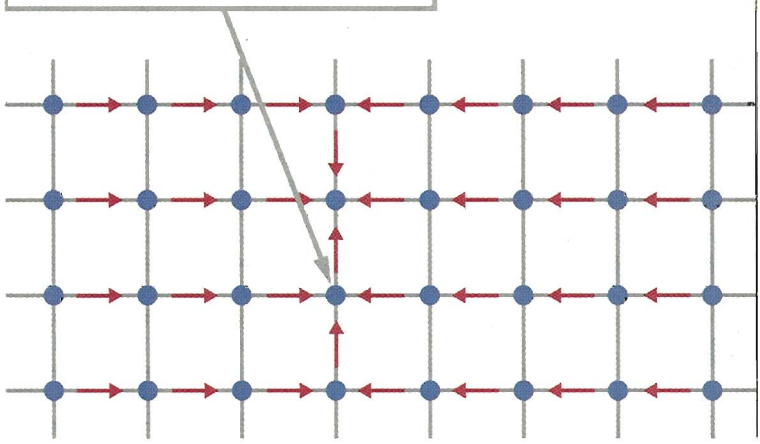
The 2MB memory on each node contains the program describing the theory the scientists wish to simulate. It comprises four 4Mb chips, each with 16 data ports. A floating-point processor was chosen because of the very large numerical range of the calculations. The one picked by the researchers was a digital signal processor (DSP), the TMS320C31–50, from Texas Instruments Inc., Dallas.

The gate array was the only component custom-designed by the researchers. It manages the memory, doing both error checking and correcting. It also manages the communications between the nodes. To receive data, the address where the data



A single node of the Riken-BNL supercomputer [left] comprises a digital signal processor, a 2MB dynamic RAM, and a custom-designed gate array that manages the memory and the communications between nodes. The nodes of the supercomputer are connected in a four-dimensional lattice with serial communication links between a node and its eight nearest neighbors: two in each of the spatial directions—*x*, *y*, and *z*—and two in time.

The arrangement of nodes [blue dots] in a two-dimensional plane of the lattice is shown below. An example of data motion during a global integer sum is indicated by the red arrows. In this operation, a specific node adds incoming serial data from off-node directions and passes the sum on to another node.



from another node is to be stored is loaded by the DSP into a memory-mapped register of the gate array, which then carries out the transfer automatically. The gate array has eight direct-memory access engines that simultaneously manage eight transfers in the eight different directions in which the node can communicate [figure, top]. The links are bidirectional and communication is serial. "So at any given moment, the gate array can be managing either the sending or receiving of eight separate data streams at a rate of 50 MHz," said Christ. The gate array also contains eight 8-word buffers so that each communication port can store the incoming data before writing it to memory.

The daughterboards, in turn, plug into motherboards measuring 356 mm by 508 mm. Each motherboard contains 63 daughterboards mounted on in-line modules and has a 64th node soldered directly onto it. This node is slightly different from the others. "We needed one node to control the functions on the motherboard. This node has a lot of address wires and data wires that go to other components on this larger host board," explained Christ.

The nodes on the motherboard are connected in a 4-by-4-by-2-by-2 mesh. An electrically programmable ROM on the motherboard stores boot code; and a small computer systems interface (SCSI) permits direct communication with the host computer, a Sun Microsystems workstation. "The motherboard is the smallest part of the system that is fully functional," said Christ. "We can load a program and do a calculation on a single motherboard."

The next level of integration is the crate, which contains eight motherboards. The crate's backplane supplies power and clock signals to the motherboards. A single crate can be placed in an air-cooled cabinet, or two of them can be mounted in a two-crate, water-cooled cabinet. The computer at Brookhaven National Laboratory contains 24 crates, while the one at Columbia has 16.

Communication is the key

A great deal of planning went into the computer's communication networks. During the physical calculations, each node talks only to its nearest neighbors. But this is rather inefficient for other situa-