from *SIAM News*, Volume 36, Number 2, March 2003

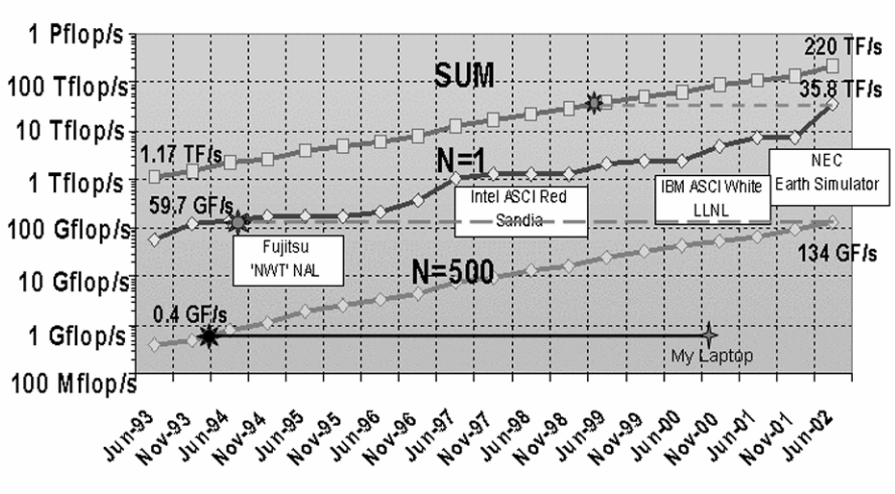
SC2002: A Terable Time for Supercomputing

By Barry A. Cipra

For anyone who's been Rip Van Winkling the last few years, teraflops—a trillion floating-point operations per second—has taken over as the standard unit of supercomputing performance. Top500.org, which rates the performance of computers on a LINPACK dense matrix calculation, announced its most recent list at Supercomputing 2002, the 15th annual supercomputing conference, held in Baltimore, November 19–22. The top 47 machines on the list all run at teraflops rates; the *slowest* runs at 195.8 gigaflops.

In first place, at 35.86 teraflops on the LINPACK benchmark, is the Earth Simulator, a Japanese machine built by NEC Corporation that began operating last March. That's more than five times faster than the two runners-up, a pair of ASCI Q's built by Hewlett–Packard and installed at Los Alamos National Laboratory. The identical twins each bench-marked at 7.727 teraflops.

The 7000-plus attendees at SC2002 heard talks on the state of the art in supercomputing technology. The meeting also included a round of awards to researchers who have developed algorithms and applications that harness the power of these fast machines.



Performance development, released with the most recent (November 2002) TOP500 list.

Paul Bunyan's Pocket Calculator

By one measure, the Earth Simulator is a throwback to olden days, when computers occupied entire rooms: The Earth Simulator takes up a whole building. Its vector architecture and custom-built processors are also throwbacks of a sort. Supercomputing in the U.S., which has long led the field, has moved mainly toward the use of commercial processors, with distributed, or "grid," computing the latest twist. The Earth Simulator reverts to the paradigm of the 1970s, when Cray was cranking out the first supercomputers. The Japanese machine has 5120 processors, each with a peak performance of 8 gigaflops, giving the Earth Simulator a theoretical peak performance of 40.96 teraflops. The processors are configured in 640 nodes of eight processors each, with 10 terabytes of main memory (16 gigabytes of shared memory per node), surrounded by 700 terabytes of disk space.

The whole works fills up the third floor of a three-story structure, 17 meters tall and 50 × 65 meters on the sides. The floor below houses the computer's copper cables, some 1800 miles of them, looking for all the world like the root system of an electronic forest. The floor below that contains the all-important air conditioning unit, needed to placate chips that produce 170 watts apiece.

The computer has its own power plant. The people who run and use it are housed in a separate building, connected to the computer's quarters by a glassed-in bridge—the only entry point for maintenance. (In photographs, the arrangement looks like a gigantic muffler and tailpipe.) The Earth Simulator is also protected by lightning rods above and a seismic isolation system below—apropos for a machine whose main mission, as reflected in its name, is to run high-resolution models in atmospheric, oceanographic, and solid earth science.

In a plenary talk at SC2002, Tetsuya Sato, director of the Earth Simulator Center, in Yokohama, described the technical innovations behind the machine. Among them are the micro-miniaturization of processors, switches, and main memory units. Each processor is a 2-centimeter square. Sixteen of them (two nodes), with all the attendant power and cooling paraphernalia, occupy a cabinet the size of a closet (1 × 2 × 1.4 meters, to be precise). The switching network is designed for high-speed data transfer, with a top rate of 10 terabytes per second.

Sato also demonstrated some of the supercomputer's applications, including a detailed simulation of a typhoon. The computational power of the Earth Simulator enables earth scientists to take a "holistic" approach, combining micro-, meso-, and macro-scales, he said. In addition to earth sciences, the Earth Simulator is slated for modeling fusion reactors, automobile and airplane aerodynamics, and computational chemistry aimed at drug design and protein studies. Describing the Earth Simulator as a bargain at $350 million, Sato asserted that it will quickly pay for itself by contributing to dramatic cost reductions in the R&D phase of manufacturing, not to mention potential savings to a society forewarned by predictions of weather disasters and—although this remains a (vector) pipedream for the successor of the Earth Simulator—earthquakes.

Not only did the Earth Simulator top the Top500 list, applications run on it won three of the five Gordon Bell prizes awarded at SC2002. Founded in 1987 by the eponymous parallel-processing pioneer, these prizes are the Oscars of supercomputing. They're also a telling measure of progress in the field: The very first Bell prize recognized work done at Sandia National Laboratories on a 1024-processor NCUBE whose processors ran at a then-blazing, now snailish 80 kiloflops.

The Earth Simulator's atmospheric general circulation model, AFES ("AGCM For the Earth Simulator"), earned its developers, led by Satoru Shingu, the prize for peak sustained performance, running at 26.58 teraflops. AFES predicts such variables as temperature, pressure, wind patterns, and humidity. The model is based on global hydrostatic equations on the sphere. Its horizontal resolution is on the order of 10 kilometers: The model uses a grid with 3840 equally spaced lines of longitude and 1920 lines of latitude placed at the Gaussian quadrature points (in essence, at the zeros of the 1920th Legendre polynomial). There are 96 layers in the vertical direction, for a total of 707,788,800 grid points.

This is one or two orders of magnitude finer than most other climate models, which typically use only a dozen or so vertical layers and rarely venture below a horizontal resolution of 100 kilometers. (A group headed by Philip Duffy at Lawrence Livermore National Laboratory put finishing touches on an 18-layer, 50-km model last year, and is now working on a 25-km simulation.) A fine mesh is not the be all and end all of climate modeling, of course; the quality of the physical parameterizations is crucial, and the amount of time, preferably measured in years, covered by a simulation is also important. But given a choice, modelers will always opt for high resolution over low.

Channeling the Data Glut

While supercomputing aficionados have become blasé about any prefix short of "tera" in describing computer speeds, "giga" still gets their attention when attached to "bits per second," as evidenced by the winners of the SC2002 Bandwidth Challenge. The winner of the "Highest Performing Application" was a multi-institution team led by John Shalf of the Lawrence Berkeley National Laboratory. They had carried out a distributed simulation of gravitational waves produced by a collision of black holes, using computers at seven locations in the U.S., the Netherlands, Poland, and the Czech Republic. The computation was coordinated by the software packages Cactus, Globus, and Visapult. With a peak transfer rate of 16.8 gigabits per second, it won handily.

A Japanese entry, the "Data Reservoir," won the award for "Most Efficient Use of Available Bandwidth," running at a peak rate of 585 megabytes per second. Developed at the University of Tokyo, the Data Reservoir is a general-purpose file-sharing facility, intended to be scalable with respect to both network bandwidth and file size. To date, it has been used mainly for transferring satellite data from the Institute for Space and Astronomical Science to the University of Tokyo, some 25 miles away.

The Bandwidth Challenge award for "Best Use of Emerging Infrastructure" went to Project DataSpace, a collaboration of researchers in Chicago, Ottawa, and Amsterdam. According to Robert Grossman, director of the National Center for Data Mining at the University of Illinois at Chicago, two streams of data—one arriving over SURFnet from the Dutch National Supercomputing Facility, SARA, at 2.8 gigabits per second and the other over Canada's CA*net[4] from a cluster of computers at Canada's advanced Internet development organization, Canarie, at 2 gigabits per second—were merged in a "lambda join" at UIC's optical network facility, StarLight. ("Lambda" is the up-and-coming buzzword for networking technology based on photons rather than electrons.)

For Grossman, the most significant thing about bandwidth lies in the implications for data mining. "With lambda joins, it is now practical to look for correlations in data even if the data is scattered around the world," he says.

That capability is likely to become increasingly important, as researchers in such data-deluged areas as bioinformatics, atmospheric and geo-science, and astronomy (not to mention the latest federally funded cybercraze, homeland defense) seek to capitalize on each other's work. Grossman and colleagues at the NCDM are developing "data webs": Web-based protocols and standards for combining and analyzing data from many remote sources. The idea is for researchers anywhere in the world to be able to use data from anywhere else in the world. In a clever bit of prefix play, Grossman calls this "terra-wide" computing.—*B.C.*

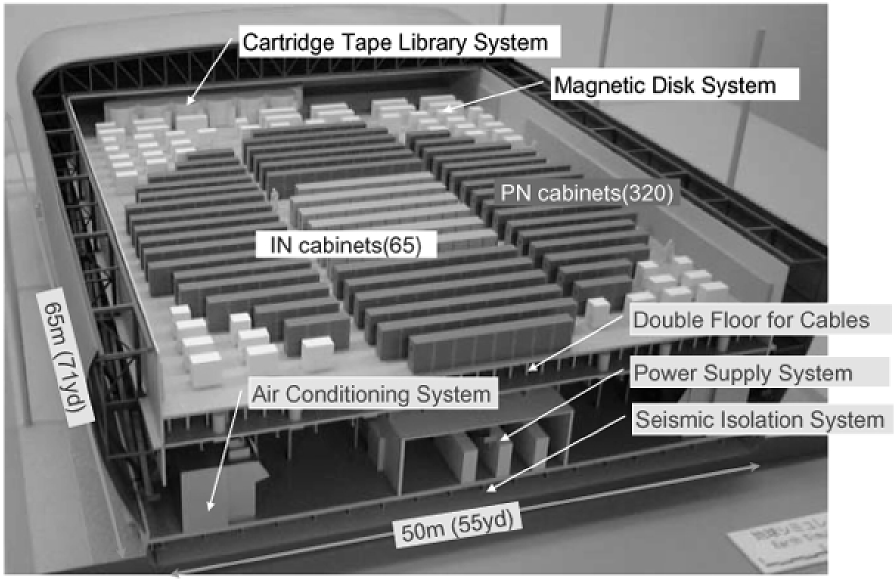
The AFES model's level of detail, its creators explained at SC2002, is near the limit of validity for the hydrostatic approximation. To go any further—say to the 1-km resolution required to model such things as convection cells in thunderstorms—will require switching to nonhydrostatic equations. But there's no need for an overhaul yet. Despite the Earth Simulator's awesome power, covering the entire globe with a grid point every kilometer is still impractical. It will take another revolution or two in supercomputing technology before researchers can hope to deploy models with meshes that fine.

As it is, most of the work in AFES goes into computing Legendre transforms. The model is technically known as T1279L96; T1279 stands for triangular truncation with a maximum wavenumber of 1279, which is the integer part of (3840-1)/3, and L96 stands for the 96 vertical layers. (In this terminology, the Lawrence Livermore model is described as T239L18.) Notice that 3840 is a multiple of 640, the number of nodes of the Earth Simulator. The record-setting, 26.58-teraflops computation is a 10-timestep (5-minute) simulation that uses all 5120 processors. In a full one-day simulation (2880 timesteps), the sustained performance was 23.93 teraflops. The 5-minute calculation took less than 5 seconds (4.651, to be precise); when run on only 80 processors (the minimum number for this particular model), it took just under 4 minutes (238.037 seconds). The speedup factor of 51.18 represents a parallel efficiency of 80%.

The other two Gordon Bells rung up by researchers working on the Earth Simulator were for a 16.4-teraflops direct numerical simulation of turbulence and a 14.9-teraflops three-dimensional plasma simulation. The turbulence computation, carried out by Mitsuo Yokokawa and colleagues, won in the "special accomplishments" category. It used a spectral method, for which the fast Fourier transform dominates the calculation. The sustained speed of 16.4-teraflops was achieved on a 2048^3 grid in single-precision arithmetic. The researchers also ran double-precision simulations on the 2048^3 grid and single-precision simulations on a 4096^3 grid.

The plasma simulation, developed by Hitoshi Sakagami and colleagues, won the Bell prize for language; it was written in High Performance Fortran, a parallel programming language developed in the early 1990s. The simulation—a 3D computation of Rayleigh–Taylor instability in an imploding target—was done on a 2048 × 2048 × 4096 grid, parallelized over 512 nodes (i.e., 4096 processors) of the Earth Simulator. It is part of the Earth Simulator's program in fusion science.

Parallel Efforts



Artist's depiction of the Earth Simulator.

Two other Gordon Bell awards for special accomplishment went to groups from Sandia National Laboratories and the University of Illinois at Urbana–Champaign. Salinas, the Sandia entry, is a scalable software package for simulating structural and solid mechanics. Salinas computes finite element models of stress, vibration, and transient dynamics with millions, even hundreds of millions, of degrees of freedom. It has been used, for example, in a model with 12 million degrees of freedom to analyze vibrations affecting the circuit boards in a smart bomb.

The award-winning computation included an analysis of an optical shutter with 110 million degrees of freedom. The calculation, which took just under 7 minutes, ran at a sustained rate of 745 gigaflops on 3375 processors of Sandia's ASCI White. An artificial example—the static analysis of a cube under pressure—ran even faster, clocking in at 1.16 teraflops.

The Illinois entry, NAMD, is a molecular dynamics code geared for parallel processing. It simulates the dynamics of macromolecules, such as proteins or nucleic acids, along with surrounding water molecules and ions, in 1-femtosecond (10^{-15} second) timesteps, using Newton's equations and an empirical energy function. Molecular dynamics, NAMD researcher Laxmikant Kale points out, is computationally easy in the sense that it's characterized by persistent repetition on a relatively small data set; it's also extremely hard, however, because it can be done only in sequential timesteps—many millions of them, to get anything biologically meaningful.

Parallelizing the computation for thousands of processors is not a straightforward proposition. Roughly speaking, NAMD proceeds by dividing space into a number of cubes and then distributing to various processors the calculations of forces acting within cubes and between pairs of adjacent cubes. In their Gordon Bell calculation, the NAMD team simulated the dynamics of 327,000 atoms in water-surrounded ATP synthase, a key protein in the cycle of living cells. The simulation was run on the 3000-processor Lemieux Alpha cluster at the Pittsburgh Supercomputing Center. At the time, NAMD's peak performance was 789 gigaflops on 2250 processors. More recently, the program has been refined to run at 1 teraflops on all 3000 processors. NAMD is freely available from the theoretical biology group at the University of Illinois (http://www.ks.uiuc.edu/Research/namd/).

The SC2002 award for best technical paper (the Emmy of supercomputing?) went to Omar Ghattas and Volkan Akcelik of Carnegie Mellon University and George Biros of the Courant Institute of Mathematical Sciences, for their paper describing a parallel algorithm for inverse wave propagation. (The proceedings of SC2002 are available online at www.sc-conference.org/sc2002.) Designing scalable algorithms for solving inverse problems based on nonlinear partial differential equations "poses a significant challenge," the researchers say. A straightforward Newton's method suffers from ungainly matrices and problems with local minima. Ghattas and colleagues get around these difficulties with a combination of Gauss–Newton linearization and matrix-free Krylov iterations.

They have applied their algorithm to the acoustic reconstruction of a pelvic bone, solving for material properties of the bone from surface pressure readings. The computation, which took three hours on 256 processors of the Terascale Computing System at the Pittsburgh Supercomputing Center, involved more than 2 million material parameters. That's small potatoes in light of their ultimate aim: estimation of the elastic properties of the greater Los Angeles Basin from historical earthquake data and a wave propagation model.

For that, the researchers may have to await the next round of supercomputers. But the wait shouldn't be long. One glimpse of the future at SC2002 came from the U.S. Department of Energy, which announced a $290 million grant to IBM to build two machines for Lawrence Livermore National Laboratory as part of DOE's National Nuclear Security Administration's Advanced Simulation and Computing program (ASC—formerly the Accelerated Strategic Computing Initiative, or ASCI). The two machines, dubbed ASCI Purple and BlueGene/L, "promise to deliver cost-effective, tremendous capability to the Stockpile Stewardship Program's critical mission," energy secretary Spencer Abraham said in the announcement. The industry–government partnership, he said, will "help solve pressing national issues, not only involving nuclear weapons, but also in areas of homeland defense, global diseases and weather prediction."

ASCI Purple will consist of 12,544 processors configured in 196 individual computers of 64 processors each. It is slated to reach the 100-teraflops mark, or roughly two and a half Earth Simulators. BlueGene/L, a spinoff of IBM's research effort in biomolecular simulations, is expected to weigh in at about 360 teraflops on 131,072 processors (2^{16} two-processor nodes) running Linux. Despite its computing power, BlueGene/L will consume a mere megawatt or so of physical power. Moreover, its "footprint" will be under 2500 square feet, about the area of a medium-sized house. The two computers are due for delivery to Livermore in 2005. Even earlier, though, the (current) computing power of the Earth Simulator is slated to be matched by a $90 million, 40-teraflops Cray computer, Red Storm, scheduled for installation at Sandia in 2004.

Clusters and Grids

Monster machines are just one facet of supercomputing. The cluster concept—harnessing scores of ordinary commercial PCs to compute in concert—is another. Indeed, cluster computers of various kinds took 93 spots on the latest Top500 list, and two made the top ten. Overall, cluster computing is likely to remain a popular approach to low-budget supercomputing.

Yet another facet that's beginning to sparkle is grid computing: making use of the Internet to focus geographically diverse computers on a single problem. The National Science Foundation has put some serious money into the effort with its TeraGrid program, an alliance of five existing supercomputing sites. TeraGrid got off the ground in 2001 with a $53 million grant to four sites: the National Center for Supercomputing Applications at the University of Illinois; the San Diego Supercomputer Center at the University of California, San Diego; Argonne National Laboratory; and the Center for Advanced Computing Research at Caltech. It expanded by another $35 million last October, to include the Pittsburgh Supercomputing Center at Carnegie Mellon University and the University of Pittsburgh.

TeraGrid's goal is to make 20-teraflops computing power openly available to the scientific community. If supercomputing history is any indicator, it will probably exceed that goal—perhaps by an order of magnitude or more.

Sooner or later, even teraflops will seem slow. The next prefix on tap is "peta," with "exa," "zetta," and "yotta" waiting in the wings. Even they may not suffice, says Julian Borrill, a computational cosmologist at Lawrence Berkeley National Laboratory, who gave a plenary talk on astrophysical simulations at SC2002.

"The universe will always exceed our computing capacity," he says. In particular, cosmologists' interests range over 50 orders of magnitude. There's only one standard that will suffice, Borrill jokes: "Bring on the googol-flop!"

Barry A. Cipra is a mathematician and writer based in Northfield, Minnesota.